Vertical integration, global and modular analysis of molecular interaction networks of *Escherichia coli*

Von Der Fakultät für Lebenswissenschaften

der Technischen Universität Carolo-Wilhelmina

zu Braunschweig

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

genehmigte

D i s s e r t a t i o n

von **Bharani Kumar Pagadala Santhanam**
aus **Bhuj / Indien**

1. Referent: Professor Dr. An-Ping Zeng
2. Referent: Professor Dr. Dietmar Schomburg
eingereicht am : 07.03.2007
mündliche Prüfung (Disputation) am: 08.05.2007

Druckjahr 2007

Vorveröffentlichungen der Dissertation

Teilergebnisse aus dieser Arbeit wurden mit Genehmigung der Fakultät für Lebenswissenschaften, vertreten durch den Mentor der Arbeit, in folgenden Beiträgen vorab veröffentlicht:

Publikationen

Ma, H., Kumar, B., Ditges, U., Gunzer, F., Buer, J. & Zeng, A. P. An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs. Nucleic Acids Res. 32: 6643-6649 (2004).

Tagungsbeiträge

Kumar, B., Ma, H. & Zeng, A. P. An integrated cellular network of *Escherichia coli* and its structural analysis. Proceedings of Foundation of Systems Biology in Engineering. 7-10 (2005).

Acknowledgments

I would like to express the deepest appreciation to my mentor, Professor An-Ping Zeng, who has the attitude and the substance of a genius. During the course of my work, he continually and convincingly conveyed a spirit of adventure in regard to research and an excitement in regard to teaching.

I would like to thank my committee members, Professor Dietmar Schomburg and Professor Reinhard Hehl, for their valuable discussions and advices during the submission of this work.

In addition, I would like to thank all my colleagues from the Research Group Systems Biology and my friends working in other departments at the Helmholtz centre for Infection Research.

Last, but yet the most important, I would like to thank my girlfriend Audrey Leprince who gave me necessary love to get things done during the last toughest period of my PhD research. I would like to thank my brothers and sister who always motivated me and especially my dad, without whose guidance and persistent help, it would have never been possible to reach this stage in my life.

Table of contents

List of figures

List of tables

Abstract

Phenotypical characteristics of cells often arise from interactions between genes, proteins and metabolites. For a complete understanding of cellular processes and their regulations it is necessary to vertically integrate the molecular networks into an interactome and understand its global structure. In this thesis,, an integrated molecular network (IMN) of *Escherichia coli* was reconstructed which comprises metabolic reactions, metabolite-protein interactions (MPI) and transcriptional regulation data. Three fundamental aspects of cellular processes were studied: (i) feedback regulation of gene expression, (ii) network motifs and (iii) global organization. Intriguingly, this work found that feedback regulation of gene expression in *E. coli* is mediated by MPIs and 69 such feedback loops (FBLs) were identified. Motif studies identified the FBL as a significant pattern and detected 12 other three-node motifs comprising five composite motifs. Connectivity analysis discovered the existence of bow-tie architecture and motif analysis in the bow-tie components revealed that 77% of them interconnect to form the giant strong component which is the backbone of the bow-tie.

Further in this work, cluster and modular analyses were performed on the integrated molecular network of *E. coli* constructed from diverse collection of datasets involving metabolic reactions, metabolite protein interactions and transcriptional regulation. Modularity was used as the parameter of an appropriate, fast and robust method for clustering such a heterogeneous molecular circuitry of interactions. This work revealed that clustering this complex network significantly grouped together genes of known similar function in well-defined physiologically related modules. Identification of network motifs and correlating them with the modules of highly connected nodes may define their potential functional role. To this end, twelve highly significant three-node network motifs among which four are composite network motifs comprising multiple types of interactions were detected and analyzed. Distribution analysis of these motifs within and between the various functional modules supported the fact that these motifs represent basic patterns of regulation and organization of genes into modules.

This thesis illustrates the potential of data integration of molecular networks to detect the feedback interactions in regulatory networks and its global analysis for better understanding cellular processes and their regulation. Moreover this work also presents a basic framework for detecting functional modules and their interaction with various motifs in an integrated *E.coli* system.

Abstract – German Version

Phenotypische Eigenschaften von Zellen entstehen häufig aus Wechselwirkungen zwischen Genen, Proteinen und Metaboliten. Für ein ganzheitliches Verstehen von Zellprozessen und ihrer Regulation ist es notwendig, die molekularen Netzwerke vertikal in ein Interactom zu integrieren und seine globale Struktur zu verstehen. In dieser Arbeit wurde ein integriertes molekulares Netzwerk (IMN) von *Escherichia coli* modelliert, dass aus den metabolischen Reaktionen, Metabolit-Protein-Wechselwirkungen (MPI) und den transkriptional-regulatorischen Elementen bestand.

Drei grundsätzliche Aspekte von Zellprozessen wurden untersucht: (i) Feedback-Regulierung der Genexpression, (ii) Netzwerkmotive und (iii) globale Organisation. Diese Arbeit lieferte faszinierende Ergebnisse: Es konnte aufgezeigt werden, dass die Feedback-Regulierung der Genexpression in *E. coli* durch MPIs vermittelt wird und 69 solcher Feedback-Schleifen (FBLs) identifiziert werden konnten. Motiv-Untersuchungen identifizierten die FBLs als ein bedeutendes Muster und entdeckten 12 andere Drei-Knoten-Motive, die fünf zerlegbare Motive umfassen. Konnektivitätsanalysen zeigten die Existenz der Bow-tie-Struktur auf und Motivanalyse der Bow-tie-Komponenten offenbarte, dass 77 % davon das GSC (giant strong component) bilden, welches das Rückgrat des Bow-tie darstellt.

Weiterhin wurden Cluster- und Modularanalysen im integrierten-molekularen Netwerk von *E. coli* durchgeführt, die auf diversen Sammlungen von Daten beruhten, die metabolische Reaktionen, Metabolit-Protein-Wechselwirkungen und transkriptionelle Regulierung beinhalteten. Modularität wurde als Parameter einer geeigneten, schnellen und robusten Methode zur Clusterung solcher heterogenen molekularen Schaltung von Wechselwirkungen genutzt. Diese Arbeit zeigte, dass die Clusterung dieses komplexen Netzwerkes Gene bekannter ähnlicher Funktion in wohl-definierten physiologisch verwandten Modulen signifikant gruppierte. Die Identifizierung von Netzwerk-Motiven und die Korrelation dieser mit Modulen hochverzweigter Knoten mag ihre potentielle funktionelle Rolle definieren. Zu diesem Zweck wurden zwölf hochsignifikante 3-

Knoten-Motive, von denen vier zusammengesetzte Netzwerkmotive multiple Typen von Interaktionen darstellen, entdeckt und analysiert. Verteilungsanalyse dieser Motive innerhalb und zwischen verschiedenen funktionellen Modulen unterstützte die Tatsache, dass diese Motive Grundmuster der Regulation und Organisation von Genen in Modulen darstellen.

Diese These illustriert das Potential der Datenintegrierung molekularer Netzwerke zur Entdeckung von Feedback-Interaktionen in regulatorischen Netzwerken und seiner globalen Analyse zur besseren Erkenntnis zellulärer Prozesse und ihrer Regulierung. Darüberhinaus zeigt diese Arbeit einen Grundrahmen für die Entdeckung funktioneller Module und ihrer Wechselwirkungen mit verschiedenen Motiven in einem integrierten System von *E. coli* auf.

Chapter 1
Introduction

1.1 Systems Biology

One of the most widely discussed and emergent fields in life sciences is systems biology. Systems biology (Bolouri and Davidson 2002; Lee *et al.* 2002) is an approach to explaining and predicting complex cellular and physiological phenomena of living organisms in terms of underlying physical and chemical processes and accompanying feedback regulations at molecular, cellular, tissue, or whole organ levels. This systems biology approach will combine mathematical modeling and simulation to complement the traditional empirical and experimental approach of biological and biomedical researches. Mathematical modeling and simulation may range from the molecular scale to organ scale models. These models and simulations will be driven by experimental data and will generate specific, explicitly testable predictions that should enable refinement of the models in response to experimental validation. This iterative development of models and experiments is a critical feature of the systems biology research.

Many new ideas are introduced with the advent of the human genome project (http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml) by scientists to transform technology-driven data acquisitions into explanatory accounts. Systems biology is one common area where many of these ideas are assembled to obtain a fundamental, comprehensive and systematic understanding of life. Apart from transforming biological understanding and practice, it has vital effects on other fields such as engineering, mathematics, physics and social science through its methods and concepts.

Systems biology can be classified into two categories, namely: pragmatic systems biology and systems-theoretic biology (O' Malley and Dupre´ 2005). Pragmatic systems biologists see a system as an interacting collection of molecules and other components. More theoretically driven systems biologists see systems as special levels of biological organization. This distinction has implications for modeling practice (Auffray *et al.* 2003). Pragmatic systems biologists normally model the interaction of molecules from the bottom-up. In bottom-up design, first the individual parts of the system are specified in great detail. The parts are then linked together to form larger components, which are in turn linked until a complete system is formed. This strategy often resembles a "seed" model, whereby the beginnings are small, but eventually grow in complexity and completeness. Systems-theoretic biologists, on the other hand, tend to model systems from the top-down. In the top-down model an overview of the system is formulated, without going into detail for any part of it. Each part of the system is then refined by designing it in more detail. Each new part may then be refined again, defining it in yet more detail until the entire specification is detailed enough to validate the model. The top-down model is often designed with the assistance of "dark boxes" that make it easier to bring to fulfillment but insufficient and irrelevant in understanding the elementary mechanisms. For both of them, the simple diagrammatic models that are routinely encountered in biology are insufficient for modeling systems. Both groups believe that a system modeling has to be mathematical in order to capture the complexities of higher-level biological organization.

Modeling is one of the important aspects of systems biology which are designed in such a way so as to generate information and transform datasets into biological perception. Mental models and mathematical models are few examples which are widely used in molecular biology. But the modeling process adapted by the systems biology proves its adequacy to explain complex systems and the various interactions involved in it between several different molecules like genes, proteins and metabolites. There are three modeling processes which are well classified in the systems biology: bottom-up, top-down and middle-out (Shen-Orr *et al.* 2002; O' Malley and Dupre´ 2005). Bottom-up and top-down modeling contradict each other. Hence it's always better to start somewhere in between these two models since both of them prove to have severe practical problems (Auffray *et al.* 2003). Pragmatic systems biologists use bottom-up modeling and systems theoretic biologists, in contrast, use top-down approach.

The major objective of systems biology is to integrate biological data at different levels. Pragmatic systems biologists think that the availability of more data and analysis is always useful whereas the systems-theoretic biologists believe that the only way systems biology could be developed is by framing the primary research questions. Hence it may prove potential in integrating these two aspects of systems biology into one. The integration of data from different sources provides an effective means to understand the complex cellular regulations. For example, in order to get a better understanding of the molecular mechanisms for diseases, metabolic and regulatory biochemical pathways must be inferred together. Lots of data from all kinds of sources like public and proprietary databases; sequence analysis, clinical findings; curated and raw data, are virtually entirely unintegrated. Moreover, because different experimental technologies provide different insights into a system, the integration of multiple data types offers the greatest information about a particular cellular process. Life Science data integration is one of the most challenging problems facing systems biology today. A number of techniques, approaches, and products are available to help scientists tackle this increasingly complex issue which includes application level integration and data level integration. The various tools required to use effectively the integrated data includes data browsers, query tools, visualization tools and data mining tools. Data browsers help users understand what is contained in the integrated data source and the Browser should lead users to an intuitive

query interface. Query tools help users ask meaningful biological questions across multiple domains and transfer the integrated data to a visualization tool for complex analysis. Visualization tools are required to help users sort through large volumes of integrated data, finding patterns and trends that would otherwise go unnoticed. Data mining tools are required by advanced users to automatically and intelligently search the integrated database to find ways to understand the data, predict future outcomes from it, and extract knowledge leading to new discoveries.

1.2 Metabolic Network

1.2.1 Metabolic Network Reconstruction

It is now possible to reconstruct the network of biochemical reactions in many organisms, from bacteria to human due to the sequencing of complete genomes. The first step is the reconstruction of the network solely based on the gene annotation information which are called genome-based network. This step can be automated and thus the genome-based network can be easily updated with the new annotation information in the databases. It is very important to integrate information from different databases to get a more complete enzyme gene list for the reconstruction. The enzyme information can be obtained from different databases. Metabolic networks are powerful tools, for studying and modeling metabolism. By correlating the genome with molecular physiology, metabolic network reconstruction and simulation provides a better understanding of the molecular mechanisms of any particular organism. A reconstruction breaks down metabolism pathways into their respective reactions and enzymes, and analyzes them within the perspective of the entire network or in other words, it collects all of the relevant metabolic information of an organism and then compiles them in a way that helps performing various types of analyses. Several gene/enzyme/reaction/pathway databases that are useful for metabolic reconstruction are listed below:

1. **Kyoto Encyclopedia of Genes and Genomes (KEGG)**. It is a "biological systems" database integrating both molecular building block information and higher-level systemic

information (Kanehisa and Goto 2000). Four main constituent databases of KEGG are (Kanehisa *et al.* 2005):

(i) **KEGG PATHWAY**: Manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks for metabolism, other cellular processes, and human diseases.

(ii) **KEGG BRITE**: Functional hierarchies and binary relations of KEGG objects, including genes, proteins, compounds, reactions, drugs, diseases, cells and organisms.

(iii) **KEGG GENES**: Gene catalogs of all complete genomes and some partial genomes with ortholog annotation (KO assignment), enabling KEGG PATHWAY mapping and BRITE mapping.

(iv) **KEGG LIGAND**: A composite database of chemical substances and reactions representing our knowledge on the chemical repertoire of biological systems and environments.

2. **Gene DataBase (GeneDB).** The GeneDB project is a core part of the Sanger Institute Pathogen Sequencing Unit's (PSU) activities. Its primary goals are:

- to provide reliable access to the latest sequence data and annotation/curation for the whole range of organisms sequenced by the PSU.
- to develop the website and other tools to aid the community in accessing and obtaining the maximum value from this data.

GeneDB currently provides access to 37 genomes, from various stages of the sequencing curation pipeline, from early access to partial genomes with automatic annotation through to complete genomes with extensive manual curation. (Details correct as of May 2006)

3. **BioCyc.** It is a collection of 261 Pathway/Genome Databases (Karp *et al.* 2005). Each Pathway/Genome Database in the BioCyc collection describes the genome and metabolic pathways of a single organism. The BioCyc databases are divided into three tiers, based on their quality.

3.1 **BioCyc Tier 1**: Intensively Curated Databases

(i) **EcoCyc**: It is a scientific database and a comprehensive source of information on the biology of the prototypical model organism *Escherichia coli* K-12 MG1655 (Keseler *et*

al. 2005). Through ongoing manual curation, extensive information such as summary comments, regulatory information, literature citations and evidence types has been extracted from 8862 publications. (As of May 2005)
(ii) **MetaCyc**: It describes enzymes and metabolic pathways for more than 300 organisms (Krieger *et al.* 2004). MetaCyc does not seek to model the complete metabolic network of any one organism, but to provide a comprehensive collection of experimentally elucidated metabolic pathways.

3.2 **Biocyc Tier 2:** These DBs were computationally generated by the PathoLogic program, and have undergone moderate amounts of review and updating. There are 13 DBs in Tier 2.

3.3 **Biocyc Tier 3:** The 246 Pathway/Genome Databases (PGDBs) in Tier 3 were generated by the PathoLogic program, which was used to predict their metabolic pathways and their operons (for bacteria only).

4. **Pathway Tools.** The Pathway Tools software is a bioinformatics software system for pathway analysis of genomes, and for creating Pathway/Genome Databases (PGDBs) such as Ecocyc (Karp *et al.* 2002b). The Pathway Tools software has three components:
(i) **PathoLogic**: Creates a new PGDB containing the predicted metabolic pathways of an organism, given a Genbank entry as input. Capabilities include:

- Predict metabolic pathways.
- Predict which genes code for missing enzymes in metabolic pathways.
- Predict operons.

(ii) **Pathway/Genome Navigator**: Supports query, visualization, and analysis of PGDBs. Capabilities include:

- Display pages for genes, proteins, operons, reactions, small molecules, and pathways.

- Genome browser including user-defined tracks. Print genome as a large poster.
- One page display of full metabolic map and transporter complement. May be zoomed, painted with omics datasets for analysis, and printed as a large poster.
- Publish PGDB on ones web site.

(iii) **Pathway/Genome Editors**: Provide interactive editing capabilities for PGDBs.

5. **ENZYME.** It is a repository of information relative to the nomenclature of enzymes (Bairoch 2000). It contains the following data for each type of characterized enzyme for which an EC number has been provided:

(i) EC number
(ii) Recommended name
(iii) Alternative names (if any)
(iv) Catalytic activity
(v) Cofactors (if any)
(vi) Pointers to the Swiss-Prot entry(s) that correspond to the enzyme (if any)
(vii) Pointers to disease(s) associated with a deficiency of the enzyme (if any)

6. **BRENDA.** It is the main collection of enzyme functional data available to the scientific community (Schomburg *et al.* 2002a; Schomburg *et al.* 2002b; Schomburg *et al.* 2004). It is available free of charge for academic, non-profit users via the internet (**www.brenda.uni-koeln.de**). It covers broad range of enzymes and frequently enzymes with very different properties are included under the same EC number. Data on enzyme function are extracted directly from the primary literature and then formal and consistency checks are done by computer programs. The data collection is being developed into a metabolic network information system with links to enzyme expression and regulation information.

7. **PUBMED.** PubMed, available via the NCBI Entrez retrieval system provides access to citations from biomedical literature. PubMed provides access to bibliographic information that includes MEDLINE, OLDMEDLINE, as well as:

(i) The out-of-scope citations (e.g., articles on plate tectonics or astrophysics) from certain MEDLINE journals, primarily general science and chemistry journals, for which the life sciences articles are indexed for MEDLINE.

(ii) Citations that precede the date that a journal was selected for MEDLINE indexing.

(iii) Some additional life science journals that submit full text to PubMedCentral and receive a qualitative review by NLM.

Several inconsistencies exist between the above mentioned databases and published literature sources regarding the metabolic information of an organism. A systematic verification is necessary to check for consistency and accuracy after the initial reconstruction of the metabolic network. New reactions not present in the databases can be added through literature search. By this means, a reconstruction combines the relevant metabolic and genomic information of an organism.

1.2.2 Metabolic Network Analysis

A metabolic network can be represented as a stoichiometric matrix where the rows represent the compounds of the reactions, while the columns of the matrix correspond to the reactions themselves. Stoichiometry is a quantitative relationship between substrate(s) and product(s) of a chemical reaction. For the functional analysis of metabolic networks, recent research has centered on three approaches; namely elementary mode analysis, extreme pathways and metabolic balance analysis (Papin *et al.* 2004).

(i) **Elementary Mode Analysis**: An elementary mode is a minimal set of enzymes that could operate at steady state, with the enzymes weighted by the relative flux they need to carry for the mode to function (Schuster *et al.* 2000). There is always a unique set of elementary modes available for a particular metabolic network (Papin *et al.* 2004). According to Stelling *et al.* (Stelling *et al.* 2002), elementary modes can be used to understand cellular objectives for the overall metabolic network. Furthermore, elementary mode analysis takes into account stoichiometrics and thermodynamics when evaluating

whether a particular metabolic route or network is feasible and likely for a set of proteins/enzymes (Schuster *et al.* 2000).

(ii) **Extreme Pathways**: Extreme pathways are convex basis vectors that consist of steady state functions of a metabolic network. A method of singular value decomposition (SVD) of extreme pathways is used in order to understand regulation of a human red blood cell metabolism (Papin *et al.* 2002; Price *et al.* 2003). For any particular metabolic network, there is always a unique set of extreme pathways available (Papin *et al.* 2004). Furthermore, a constraint-based approach is defined (Price *et al.* 2003), where through the help of constraints like mass balance and maximum reaction rates, it is possible to develop a 'solution space' where all the feasible options fall within. Then, using a kinetic model approach, a single solution that falls within the extreme pathway solution space can be determined (Price *et al.* 2003). Therefore, in the study (Price *et al.* 2003), both constraint and kinetic approaches are used to understand the human red blood cell metabolism. In conclusion, using extreme pathways, the regulatory mechanisms of a metabolic network can be studied in further detail.

(iii) **Flux Balance Analysis**: This method uses linear programming, but in contrast to elementary mode analysis and extreme pathways, only a single solution results in the end. Linear programming is usually used to obtain the maximum potential of the objective function that you are looking at, and therefore, when using flux balance analysis, a single solution is found to the optimization problem (Schuster *et al.* 2000). In a flux balance analysis approach, exchange fluxes are assigned to those metabolites that enter or leave the particular network only. Those metabolites that are consumed within the network are not assigned any exchange flux value. Also, the exchange fluxes along with the enzymes can have constraints ranging from a negative to positive value (ex: -10 to 10). Furthermore, this particular approach can accurately define if the reaction stoichiometry is in line with predictions by providing fluxes for the balanced reactions. Also, flux balance analysis can highlight the most effective and efficient pathway through the network in order to achieve a particular objective function. The enzyme that correlates to the gene that needs to be removed is giving a constraint value of 0. Then, the reaction that the particular enzyme catalyzes is completely removed from the analysis.

1.3 Transcriptional Regulatory Networks

The knowledge of biological information continues to increase and at times more information seems to preclude greater knowledge. Thousands of genes that contribute to smooth functioning of the cell are controlled by transcriptional regulatory proteins, which bind to genes and increase or decrease the rate at which the gene is transcribed. The basic unit of gene regulation consists of a transcription factor, its DNA binding site and the target gene or transcription unit it regulates. This basic unit can be elaborated to form a complex network in two ways: some genes may be regulated by more than one transcription factor, and some transcription factors may control more than one gene. The procedure in which a collection of regulatory proteins associate with genes across a genome can be described as a hardwired genomic regulatory code (Shen-Orr *et al.* 2002) and the ultimate function of which is to set up a progression of transcriptional regulatory states in space and time. For experimental manipulation and, most fundamentally, for comprehension of how transcriptional regulatory networks (TRNs) work, models are required for analysis. For this purpose, a thorough knowledge of their overall structure and the modular building blocks of which the TRN are hierarchically constructed are needed. The overall structure mainly depends on the linkage among the regulatory genes and the modular building blocks consist of basic transcriptional control processes executed by one or a few functionally linked genes (Shen-Orr *et al.* 2002; Milo *et al.* 2004).

In integrating genome-wide data on transcript abundance into a dynamic view of gene networks, recent studies have focused on abstracting the principles that underlie the architecture and causal interplay of these networks (DeRisi *et al.* 1997). In order to know how far systems biology can help us, it is necessary to find or use suitable model organisms on which the novel ways of understanding and new modeling strategies can be tested. The organism *Escherichia coli* K-12 fits best into this category since it is the best known and natural biological model. Several reference databases are available on *E.coli* K-12. For example, RegulonDB contains the curated knowledge of genetic regulation and operon organization (Salgado *et al.* 2004). The other well known repository for *E.coli* is Ecocyc. One of the significant progresses in the recent years is the unification of these

two databases to provide the users with more detailed information and avoiding confusions. Even though the amount of experimentally validated knowledge of the *E.coli* regulatory network is the largest currently available for any organism, we are left open with questions regarding gene function, regulatory mechanisms, or global integration. Regulation of gene expression in *E.coli* involves a complex network and DNA-binding transcription factors are an important component of this network since they respond to changes in the cellular environment by altering the gene expression of relevant genes. Recent research has elucidated interesting aspects of the design principles of the transcriptional regulation network (Shen-Orr *et al.* 2002; Milo *et al.* 2004), including the motifs that recur in the network and their functions, scale free characteristics (Ravasz *et al.* 2002) and others. Recent studies have also focused on the significance of the arrangement of the genes within the genome and its relationship to the regulatory network architecture (Korbel *et al.* 2004; Warren and Wolde 2004).

Besides genome sequencing, a series of large scale experimental analyses have been initiated, aiming at uncovering the functional organization of cells. In order to disentangle gene regulatory networks at the level of the whole organism, several groups started systematic global studies of gene expression and DNA-protein interactions in different conditions (Wang and Church 1992; Chuang *et al.* 1993; VanBogelen *et al.* 1996). Clearly, such time or space scale snapshots of gene expression in various conditions will be of great help in the delineation of the main regulatory pathways. As a complement to these experimental approaches, there is an increasing need for efficient theoretical tools and formal frameworks to derive regulatory structures from partial expression data (Collado-Vides *et al.* 1996; Palsson 1997; Strohman 1997).

1.4 Protein-Protein Interaction

Protein-protein interactions refer to the association of protein molecules. Many biological functions involve the formation of the protein-protein complexes. Proteins might interact for a long time to form part of a protein complex, a protein may be carrying another

protein or a protein may interact shortly with another protein just to modify it. This modification of proteins can itself change protein-protein interactions. For virtually every process in a living cell, protein-protein interactions are of central importance. Information about these interactions improves our understanding of diseases and can provide the basis for new therapeutic approaches.

There are lots of methods available to investigate protein-protein interactions. They are discussed below:

(i) **Co-immunoprecipitation**: It is considered to be the gold standard assay for protein-protein interactions, especially when it is performed with endogenous (not over expressed and not tagged) proteins. The protein of interest is isolated with a specific antibody. Interaction partners which stick to this protein are subsequently identified by western blotting. Interactions detected by this approach are considered to be real. However, this method can only verify interactions between suspected interaction partners. Thus, it is not a screening approach.

(ii) **The Yeast Two-hybrid Screen**: It investigates the interaction between artificial fusion proteins inside the nucleus of yeast. This approach can identify binding partners of a protein in an unbiased manner. However, the method has a notorious high false-positive rate which makes it necessary to verify the identified interactions by co-immunoprecipitation.

(iii) **Tandem Affinity Purification (TAP)**: It detects interactions within the correct cellular environment (e.g. in the cytosol of a mammalian cell). This is a big advantage compared to the yeast two-hybrid approach. However, the TAP tag method requires two successive steps of protein purification. Thus, it can not readily detect transient protein-protein interactions.

(iv) **Quantitative Immunoprecipitation Combined with Knock-down (QUICK)**: It relies on co-immunoprecipitation, quantitative mass spectrometry (SILAC) and RNA interference (RNAi). This method detects interactions among endogenous non-tagged proteins. Thus, it has the same high confidence as co-immunoprecipitation. However, this method also depends on the availability of suitable antibodies.

Some of the biological databases which stores the protein-protein interaction data are mentioned below:

- **APID Agile Protein Interaction Data Analyzer**: It is an interactive bioinformatics web tool to explore and analyze in unified and comparative platform main currently known information about protein–protein interactions.
- **BioGRID**: It is a public repository for protein and genetic interactions.
- **HPRD Human Protein Reference Database**: It is a (manually) curated database of human protein information with visualization tools.
- **IntAct Interaction Database**: It is a public repository for manually curated molecular interaction data from literature.
- **The MIPS Mammalian Protein-Protein Interaction Database**: It is a collection of manually curated high-quality PPI data collected from the scientific literature by expert curators

1.5 Cell Signaling

Cells must be ready to respond to essential signals in their environment. These are often chemicals in the extracellular fluid (ECF) from: (i) distant locations in a multicellular organism - endocrine signaling by hormones; (ii) nearby cells - paracrine stimulation by cytokines; or (iii) secreted by them (autocrine stimulation). Signaling molecules may trigger:

- An immediate change in the metabolism of the cell (e.g., increased glycogenolysis when a liver cell detects adrenaline);

- An immediate change in the electrical charge across the plasma membrane (e.g., the source of action potentials);

- A change in the gene expression (These responses take more time.)

Cells receive information from their environment through a class of proteins known as receptors. Notch is for example a cell surface protein that functions as a receptor. Animals have a small set of genes that code for signaling proteins that interact specifically with Notch receptors and stimulate a response in cells that express Notch on

their surface. Molecules that activate (or, in some cases, inhibit) receptors can be classified as hormones, neurotransmitters, cytokines, growth factors but all of these are called receptor ligands. The details of ligand-receptor interactions are fundamental to cell signaling.

In some cases, receptor activation caused by ligand binding to a receptor is directly coupled to the cell's response to the ligand. Often, the behavior of a chain of several interacting cell proteins is altered following receptor activation. The entire set of cell changes induced by receptor activation is called a signal transduction mechanism or pathway. Some signaling transduction pathways respond differently depending on the amount of signaling received by the cell. For instance the Hedgehog (cell signaling) activates different genes depending on the amount of Hedgehog protein present. Complex multi-component signal transduction pathways provide opportunities for feedback, signal amplification, and interactions inside one cell between multiple signals and signaling pathways.

1.6 Need for Integration

Two prominent paradigms in modern biology about which there has been much discussion are the reductionist and integrative (or systems) approaches to biological discovery and understanding (Ideker *et al.* 2001; Uetz *et al.* 2002; Barabasi and Oltvai 2004; Bateman *et al.* 2004; Yeger-Lotem *et al.* 2004). The utility of the systems approach is demonstrated when computational models are used to integrate large amounts of molecular data to interpret and predict the range of cellular phenotypes. Many biological sources are available in the internet as of date (Bateman *et al.* 2004), which stores the data regarding transcription-regulatory networks, metabolic pathways, protein-protein interactions, cellular signaling, genomes, mRNA and protein structures. An integrative approach is required in order to analysis this huge and complex amount of data in order to understand the structure and behavior of the complex intercellular web of molecular interactions that controls cell behavior (Barabasi and Oltvai 2004). The key point to understand the structure and behavior of the cell is to integrate the available data in a way

to increase our understanding of the underlying biological processes that operate inside the cell (Ideker *et al.* 2001; Uetz *et al.* 2002; Barabasi and Oltvai 2004; Yeger-Lotem *et al.* 2004).

Biological data are disseminated in many different databases. These databases have different management systems, formats and views of how to represent the data stored. Most of them are accessible by flat files or by web interfaces that allows some kind of query over it. The two main problems involved here are the difficulty in parsing the data when dealing with heterogeneous flat file formats and the inconsistency due to the absence of a unified vocabulary that causes the same information to be represented in more than one way. Fortunately, systems biology acts as a platform to improve this scenario.

1.7 Java Technology

Java technology was created as a computer programming tool in a small, secret effort called "the Green Project" at Sun Microsystems in 1991. The initial conclusion of this project was that at least one significant trend would be the convergence of digitally controlled consumer devices and computers. A device-independent programming language code-named "Oak" was the result. As it turns out, the Internet was ready for Java technology, and just in time for its initial public introduction in 1995, the green team was able to announce that the Netscape Navigator Internet browser would incorporate Java technology. Now, nearing its twelfth year, the Java platform has attracted over 5 million software developers, worldwide use in every major industry segment, and a presence in a wide range of devices, computers, and networks of any programming technology.

In fact, its versatility, efficiency, platform portability, and security have made it the ideal technology for network computing, so that today, Java powers more than 2.5 billion devices. Today, Java technology is just about everywhere in networks and devices that range from the Internet and scientific supercomputers to laptops and cell phones and credit cards etc. The reasons why the software developers choose java technology is since

it is thoroughly refined, extended, tested, and proven by an active community of over five million software developers. It allows developers to write software on one platform and run it on practically any other platform, create programs to run within a web browser and web services, develop server-side applications for online forums, stores, polls, hyper text markup language (HTML) forms processing, combine Java technology-based applications or services to create highly customized applications or services and write powerful and efficient applications for mobile phones, remote processors, low-cost consumer products, and practically any device with a digital heartbeat.

1.8 Topological Analysis of Networks

Bow-tie Structure. The development of the "Bow Tie" Theory is as a result of the most intensive research study of the web conducted by the Scientists from IBM Research and other corporate research labs. The outcome of this study proved that the web is less connected than previously thought. Broder *et al.*, identified five distinct regions of web which includes: (i) Central core, (ii) IN, (iii) OUT, (iv) Tendrils and Tubes, (v) Disconnected pages. A central core contains pages between which users can surf easily. Another large cluster, labeled 'IN', contains pages that link to the core but cannot be reached from it. A separate 'OUT' cluster consists of pages that can be reached from the core but do not link to it, such as corporate websites containing only internal links. Other groups of pages, called 'Tendrils' and 'Tubes', connect to either the in or out clusters, or both, but not to the core, whereas some pages are completely unconnected and are called ‘Disconnected’ (Fig 1). The Bow Tie Theory explained the dynamic behavior of the Web, and yielded insights into the complex organization of the Web. The discoveries help computer scientists better understand the structure of the Internet, and lead to new technologies and design advances that will speed and simplify e-business.

The bow-tie architecture is found in transcriptional and translational processes as well (Csete and Doyle 2004). On the intermediate timescale of transcriptional regulation and translation, the bow tie facilitates feedback regulation. But, probably the best studied biological network is the bacterial metabolism and bacterial metabolic networks are a

striking example of bow-tie organization. Studies revealed that only an organization such as the bow tie facilitates the type of extreme heterogeneity that allows for robust regulation, manageable genome sizes and biochemically plausible enzymes (Csete and Doyle 2004). Bow-tie structures and protocols are found throughout biology in parallel or convergent systems, as well as in homologous systems. The ubiquity of bow-tie structures in advanced technologies supports the large amount of biological evidence indicating that these structures are universal and fundamental organizing principles.

In general, power grid (where several different energy sources are used to make a universal 60-Hz AC common carrier, which in turn is widely disseminated to provide power to a large and rapidly changing variety of uses), manufacturing (where numerous raw materials, which are transformed into relatively few building blocks, which are then assembled into many different products) and money (a common carrier for the exchange of varied goods and services) implement a bow-tie protocol. Furthermore, the basic framework of bow ties is used throughout advanced technologies. Taken together, the convergent evolution in biology and developments in technology suggest that these structures and protocols are universal (Csete and Doyle 2004).

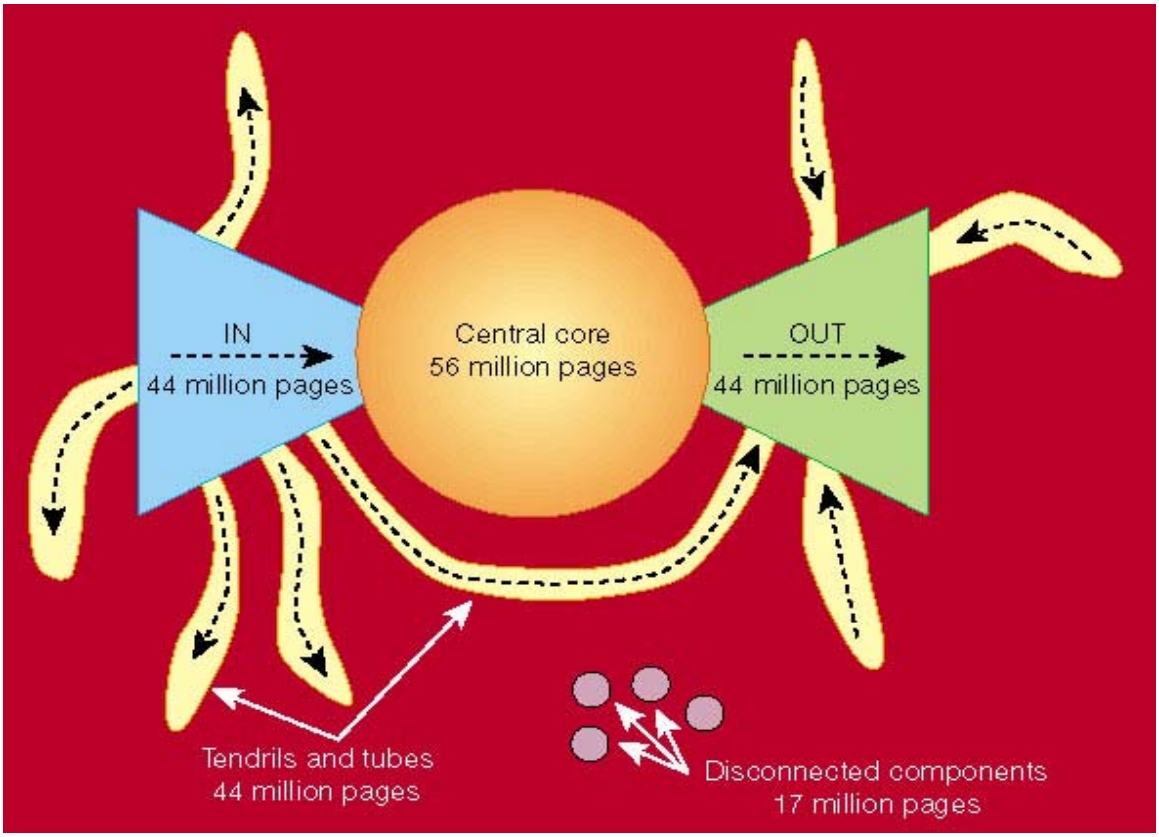


Figure 1. The bow-tie structure in the web.

1.9 Motifs

Two seminal studies (Milo *et al.* 2002; Shen-Orr *et al.* 2002) recently have shown that topological networks indeed contain statistically significant patterns indicative of biological functions. These ‘‘network motifs’’ are patterns of interconnections occurring in networks at numbers that are significantly higher than those in randomized networks that are uniformly drawn from the networks with the same degree distributions as the original networks (Milo *et al.* 2002; Milo *et al.* 2004) and the probability of the appearance of these patterns an equal or greater number of times than in the real network is lower than a cutoff value (here $P = 0.01$). The motifs found so far have been identified because they occur identically at different positions in a network. These network motifs may define universal classes of networks in that similar motifs have been found in a wide variety of networks, ranging from the World Wide Web to the electronic circuits, from the transcriptional regulatory networks of *Escherichia coli* to the neural network of *Caenorhabditis elegans*. But the motifs shared by ecological food webs were distinct

from the motifs shared by the genetic networks of *Escherichia coli* and *Saccharomyces cerevisiae* or from those found in the World Wide Web.

Recent study from (Milo *et al.* 2002) has developed an algorithm to detect network motifs and it has been applied on wide range of bidirected networks like transcription network, neuron synaptic connection network and ecological food web. This algorithm scanned each of the above mentioned networks for all possible 3 and 4-node subgraphs and found several of them. But in order to focus on the important subgraphs, the real network is compared with the randomized network and only those patterns appearing in the real network at numbers significantly higher than those in the randomized networks are selected (Fig 2).

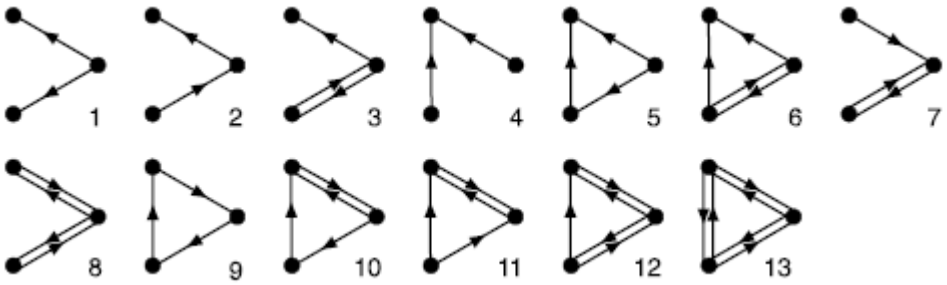


Figure 2. All 13 types of three-node connected subgraphs.

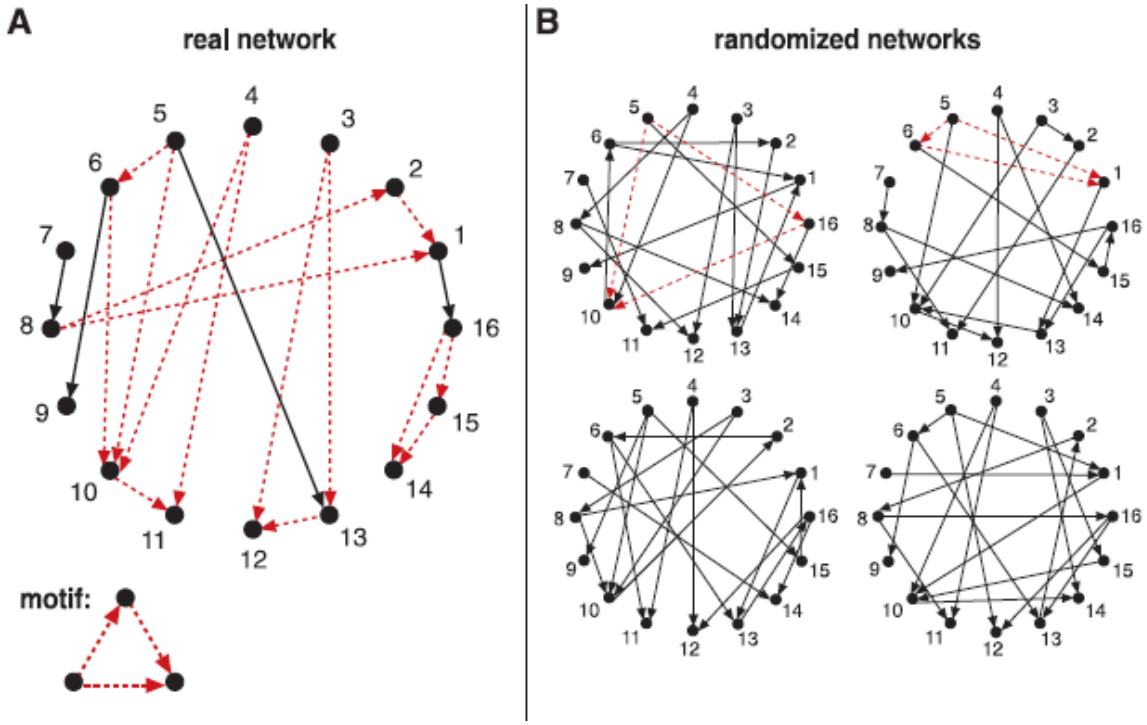


Figure 3. Schematic view of network motif detection.
Network motifs are patterns that recur much more frequently (A) in the real network than (B) in an ensemble of randomized networks. Each node in the randomized networks has the same number of incoming and outgoing edges as does the corresponding node in the real network. Red dashed lines indicate edges that participate in the feedforward loop motif, which occurs five times in the real network.

Each node in the randomized networks has the same number of incoming and outgoing edges as does the corresponding node in the real network. This makes the randomized network to have the same single node characteristics as a real network (Fig 3).

Some of the motifs found from the analysis of TRN of *Saccharomyces cerevisiae* are discussed below:

(i) **Autoregulation Motif**: An autoregulation motif consists of a regulator that binds to the promoter region of its own gene.

(ii) **Feedforward Loop**: A feedforward loop consists of a master regulator that controls a second regulator, and both regulators bind a common target gene.

(iii)**Multi-component Loop**: A multi-component loop motif consists of a regulatory circuit whose closure involves two or more factors.

(iv) **Regulator Chain**: Regulator chain motifs consist of chains of three or more regulators in which one regulator binds the promoter for a second regulator, the second binds to the promoter for a third regulator, and so forth.

(v) **Single Input Modules**: Single input modules contain a single regulator that uniquely binds a set of genes under a specific condition.

(vi) **Multi-Input Motifs:** Multi-input motifs consist of a set of regulators that bind together to a set of genes.

1.10 Cluster Analysis

The term cluster analysis (first used by Tryon, 1939) actually encompasses a number of different classification algorithms. A general question facing researchers in many areas of inquiry is how to organize observed data into meaningful structures, that is, to develop taxonomies. Clustering techniques have been applied to a wide variety of research problems. For example, in the field of medicine, clustering diseases, cures for diseases, or symptoms of diseases can lead to very useful taxonomies. In the field of psychiatry, the correct diagnosis of clusters of symptoms such as paranoia, schizophrenia, etc. is essential for successful therapy. In archeology, researchers have attempted to establish taxonomies of stone tools, funeral objects, etc. by applying cluster analytic techniques. In general, whenever one needs to classify a "mountain" of information into manageable meaningful piles, cluster analysis is of great utility.

There are many clustering methods available, and each of them may give a different grouping of a dataset. These methods are divided into *partitioning* methods and *agglomerative* methods. The most important clustering methods are described below:

1.10.1 Partitioning Methods

The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. If the

number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset.

1.10.2 Hierarchical Agglomerative Methods

The hierarchical agglomerative clustering methods are most commonly used. The construction of a hierarchical agglomerative classification can be achieved by the following general algorithm.

1. Find the 2 closest objects and merge them into a cluster
2. Find and merge the next two closest points, where a point is either an individual object or a cluster of objects.
3. If more than one cluster remains , return to step 2

Individual methods are characterized by the definition used for identification of the closest pair of points, and by the means used to describe the new cluster when two clusters are merged.

There are some general approaches to implementation of this algorithm, these being stored matrix and stored data, are discussed below:

- In the second matrix approach , an N*N matrix containing all pairwise distance values is first created, and updated as new clusters are formed. This approach has at least an O(n*n) time requirement, rising to $O(n^3)$ if a simple serial scan of dissimilarity matrix is used to identify the points which need to be fused in each agglomeration, a serious limitation for large N.
- The stored data approach required the recalculation of pairwise dissimilarity values for each of the N-1 agglomerations, and the O(N) space requirement is therefore achieved at the expense of an $O(N^3)$ time requirement.

Chapter 2
Softwares Used

2.1 Introduction

In systems biology, large number of different packages for modeling, analysis, visualization and general data manipulation are available. Several heterogeneous application components are written in diverse programming languages and run on different platforms. As of now, there are 75 different packages available for cellular simulation. But there are two major problems which the systems biology faces:

1. Each tool uses its own format, often undocumented, and the models generated and saved by this tool cannot be used by used by another tool. This acts as an obstacle for exchanging data generated by one tool to the other.
2. Many tools duplicate each other's capabilities. As a result, many of the tools provide similar functionality.

In this work, certain software tools are used to overcome these problems which could convert data from one format to other and the result generated by one of the software tools could be reused by the others.

2.2 Cluster Tool

The cluster tool version 2.0 is used in this work for the purpose of network decomposition of integrated molecular network (IMN) into functionally related modules. IN order to use this tool, the Python (http://python.org/) language and the NetworkX (http://networkx.sourceforge.net/) library need to be installed. To load the IMN in cluster tool, it should be provided in one of the supported formats such as simple interactions format (SIF) file and moreover some parameters need to be specified so that the tool can cluster accordingly, for example, the maximal or minimal number of modules to be created or let the tool select the best number of modules etc. It is available as a library called *libclust* and it is much faster than its previous versions.

2.3 Cytoscape

Cytoscape is an open-source community software project for integrating biomolecular interaction networks with high-throughput expression data and other molecular states into a unified conceptual framework (Shannon *et al.* 2003). The central organizing metaphor of Cytoscape is a network graph, with genes, proteins, and molecules represented as nodes and interactions represented as links, i.e. edges, between nodes. It runs on Linux, Mac OS X and Windows provided java is installed on these machines. Unlike other visualization software, Cytoscape has the capability to handle really large networks such as transcriptional regulation interaction or protein-protein interaction network of *Saccharomyces cerevisiae* and *Escherichia coli.* Any interaction network could be imported into Cytoscape for visualization purpose. But the main criterion is that the files should be in the simple interactions format (.sif) which is convenient for building a graph from a list of interactions. New interactions could be added an existing .sif file and moreover different interaction types could be loaded in a same file like TRIs and PPIs. One of the disadvantages of this file format is that it doesn’t posses any layout information. Hence, the next time when the same network is loaded into Cytoscape, it may generate a different layout completely different from the previous one.

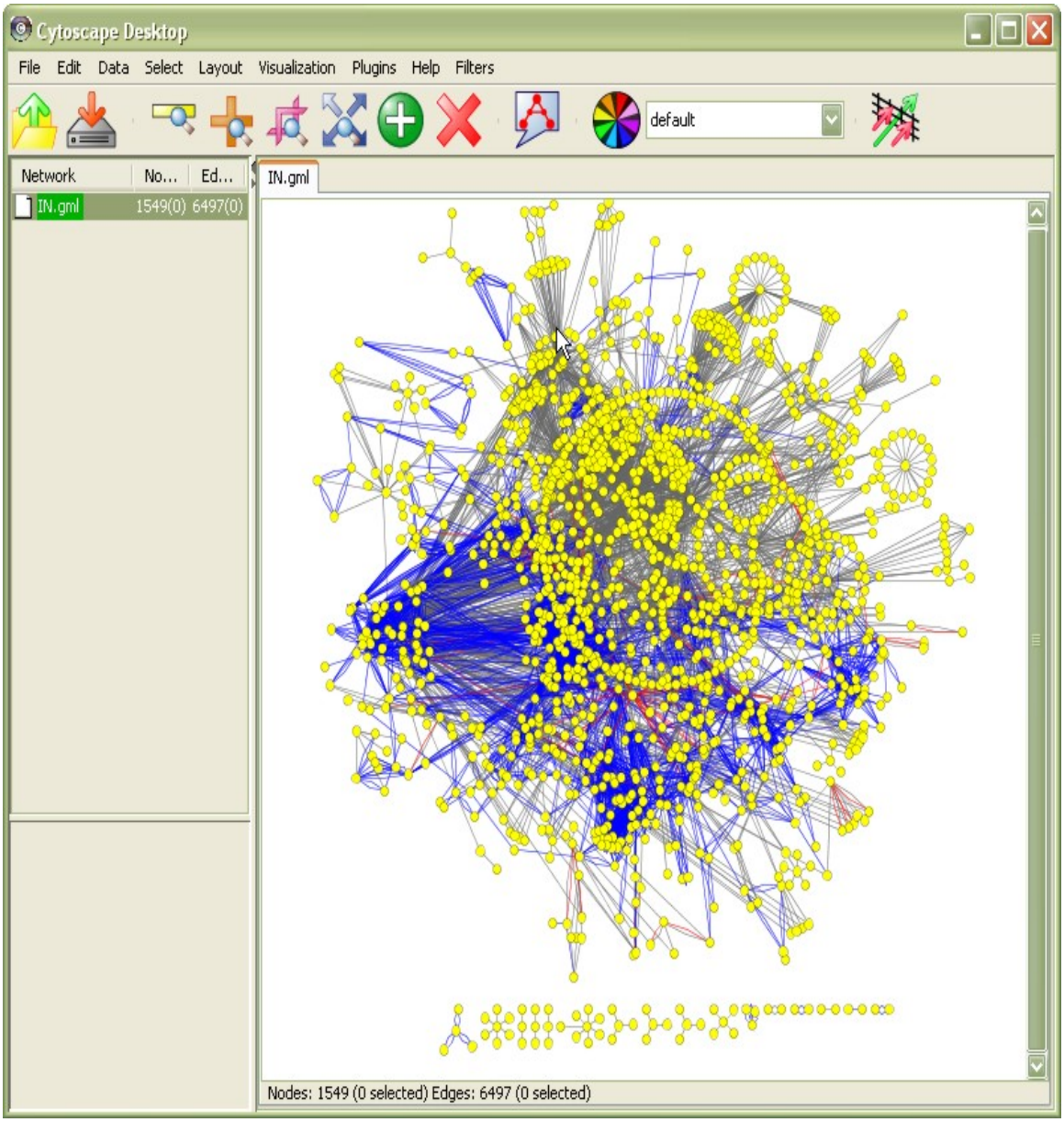


Figure 4. Cytoscape window visualizing an IMN of E.coli displaying multiple interactions in different colors.

Lines in the SIF file specify a source node, a relationship type (or edge type), and one or more target nodes:

nodeA <relationship type> nodeB

An example of SIF file: An integrated network of *E.coli* comprising multiple interaction types.

treR <transcription regulation > *treC*

treB <Metabolic reaction > *treA*

treB <Metabolite-Protein interaction> *treR*

The software reads the first line and identifies two nodes, *treR* and *treC* draws an edge between which the relationship (i.e., TRI) between the two genes. Then it moves on the next line and similarly draws an edge between the two nodes *treB* and *treA.* But in the case of the second line the relationship between the two genes (i.e., MPI) is different from the one mentioned in the first line. In this case, cytoscape can represent different interactions by drawing different edges of different colors, for example: red edge for MPI, blue edge for MR and a grey edge for TRI (Fig 4). Moreover, in order to distinguish various interaction types, cytoscape offers the user to add arbitrary node and edge information.

A node attribute file begins with the name of the attribute on the first line, and on each following line, has the name of the node, followed by an equals sign, and followed by the value of that attribute. Numbers and text strings are the most common attribute types. All values for a given attribute must have the same type. For example, the IMN of *E.coli* is clustered into functionally related modules and genes belonging to a particular functional module could be represented different from the others by loading a node attribute file as:

Functional Category.
arcA = respiration
fnr = respiration
ilvC = Leucine Biosynthesis
ilvD = Leucine Biosynthesis

By loading this node attribute file, nodes in the respiration and leucine biosynthesis modules could be represented in different colors (Fig 5). An edge attribute file has much the same structure, except that the name of the edge is the source node name, followed by the interaction type in parentheses, followed by the target node name. Along with the node and edge attributes, Cytoscape also provides visual style features which could be used to customize the visual appearance of the network. For example, one can specify a

default color and shape for all nodes or use specific line types to indicate different types of interactions.

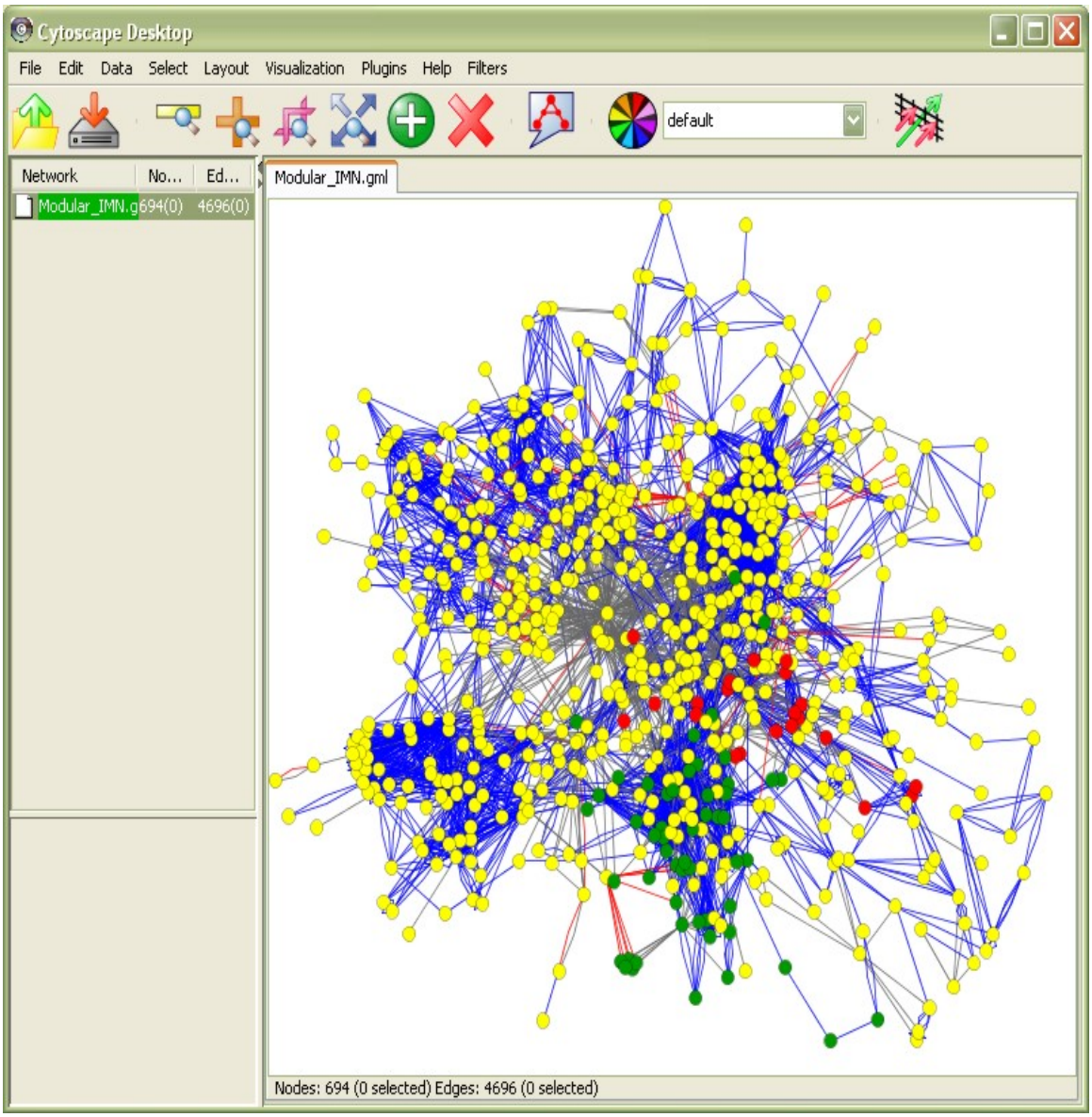


Figure 5. Cytoscape window displaying different colors for genes belonging to different modules. Genes involved in respiration is represented by green nodes and genes involved in leucine biosynthesis are represented by red nodes.

Once the network is built in from SIF format and laid out using all these attributes and visualization features, it could be saved by exporting it as graph modeling language

(GML) format. In contrast to SIF, GML is a rich graph format language supported by many other network visualization packages. For example, figure 6 shows the hierarchical layout of the extended TRN of *E.coli* (Ma *et al.* 2004c) which is saved and loaded as GML format.

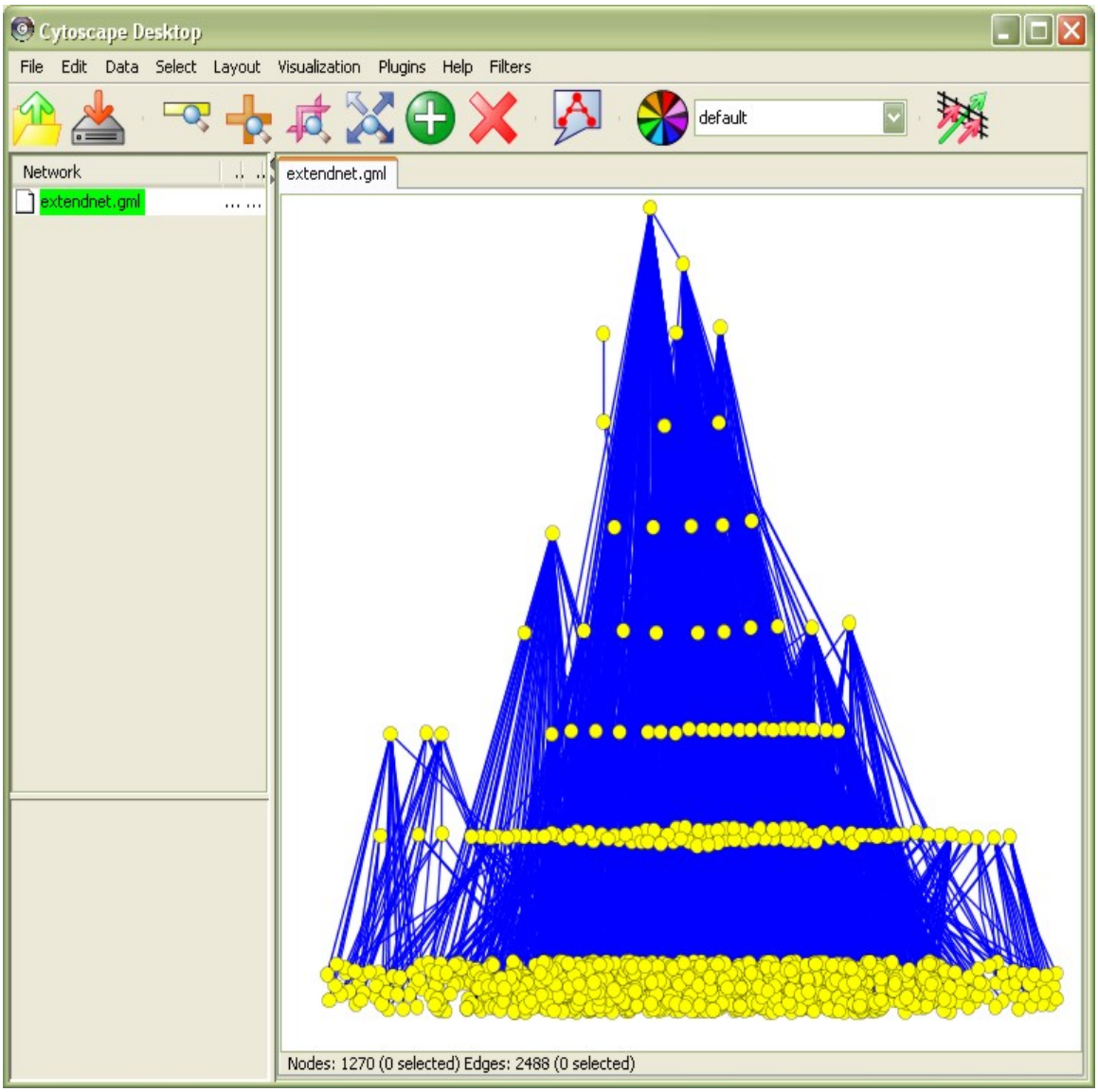


Figure 6. Cytoscape window displaying the hierarchical layout of extended TRN of E.coli which is loaded in GML format.

A sample of GML file format specification for the extended TRN:

graph [

```
node [
        label   "modB"
        id      0
        graphics [
                y       1691.6
                x       1615.4
                z       1000.0
        ]
]
node [
        label   "fruB"
        id      1
        graphics [
                y       1665.2
                x       1611.8
                z       1000.0
        ]
]
```

The GML file format specification is available at:
http://www.infosun.fmi.uni-passau.de/Graphlet/GML/
Colors and other visual attribute defined in the GML file are not currently honored by Cytoscape, only the node labels and layout information.

2.4 Java

In this work, most of the programming is done using java. For example, the relation between protein-promoter and protein–binding site relationships, which genes, promoters and binding sites are present in a particular transcriptional unit and protein–gene information respectively are all widely distributed in several files downloaded from the ecocyc database (www.ecocyc.org). In order to assemble all these data together to generate a single database file which includes all the above mentioned information, java

programming has been used to fetch the data from each of the files and linking them to the others using primary and secondary keys.

Moreover Java 2 Runtime Environment (JRE), version 1.4.2 or higher is needed in order to run Cytoscape. It can be downloaded at: http://java.sun.com/j2se/1.4.2/download.html

A new Java system property introduced in Java 1.4: java.awt.headless can be used by Cytoscape. This property allows the Java system to run without Graphics support; Cytoscape running in this mode allows users to run non-graphical analyses as batch jobs or on systems without keyboard/mouse/display capabilities, such as compute servers.

2.5 Network Conversion

The basic idea of integrating the molecular networks into a single integrated network started from the extended TRN where there was no feedback regulations noticed. But the extended TRN was in a .NET format which is being used by pajek visualization software. But, to visualize a complex network like IMN of *E.coli*, it is preferable to use software like Cytoscape rather than Pajek, since Cytoscape has more functions for visualization than Pajek. But the file format used by Cytoscape is not .NET, rather it is .SIF. Hence in order to convert the .NET file to .SIF file, the software pyNetConv is used. It can import and export multiple file formats including Pajek networks (.net) files, Pajek cluster (.clu) files, Cytoscape networks (.sif) files, Cytoscape node attributes (.na) files and GML files.

2.6 Network Motif Detection Tool

The “network motifs” are those patterns for which the probability P of appearing in a randomized network an equal or greater number of times than in the real network is lower than a cutoff value (here $P = 0.01$) (Shen-Orr *et al.* 2002). Patterns that are functionally important but not statistically significant could exist, which would be missed by our approach. In order to find out these patterns, the network motif detection tool ‘mfinder 1,1’ is used. This version of the tool can be used for:

1. Detection of larger motifs (more than 4 nodes).

2. Output of all subgraphs members (that is the names/indexes participating in each subgraph)

3. Sampling algorithm implementation which is useful for very large networks or large subgraph size.

4. Output the top 7 motifs in cases there are many motifs found (Complete motif list can be found in the output file)

Input network file format should be a simple text (.txt) format. Nodes in the network should be represented by integers. Each edge in the network should be represented by an equivalent line of the following format:

<source node> <target node> <edge weight>.

Example:

1 2 1

3 1 1

represents a network of 3 nodes with two edges:

(1->2) and (3->1)

In this version of motif analysis tool, the edge weight is ignored and should be 1 for all edges. But this tool could be used only for the interaction network comprising one single type of interaction, for example, transcriptional regulation interactions and metabolic reaction networks. Moreover it could be used for the non-directed interaction networks like protein-protein interaction network with a condition that every edge should appear only once.

In case of an integrated network which comprises multiple interaction types (Yeger-Lotem *et al.* 2004), this tool could not be used since the randomization procedure used by mfinder 1.1 is unique for single interaction type networks. Hence the tool mfinder 042r is used which can handle integrated network comprising multiple interaction types up to three different interactions (Yeger-Lotem *et al.* 2004). This tool extends the approach of (6) to generate randomized network containing multiple type edges. (Shen-Orr *et al.* 2002) generated randomized networks with the same network characteristics by preserving the node degrees (Shen-Orr *et al.* 2002). For dealing with networks with multiple types of connections we defined two terms:

1. Extended degree of a node.

2. Edge profile of nodes

The extended degree reflects the local connectivity of a node, and the edge profile provides a local measure of the relation between two nodes. The randomized networks are generated such that both the extended degree of each node and the profile of each edge in the network are retained. This tool generates randomized networks by an iterative switching of edges using the four-point-switchability condition (section 4.2.4).

2.7 Pajek

It is a program, for Windows, for analysis and visualization of large networks having some thousands or even millions of vertices (Bategelj and Mrvar 1998). It provides tools for analysis and visualization of of large networks that are already in machine-readable form, for example, collaboration networks, organic molecule in chemistry, protein-receptor interaction networks, genealogies, Internet networks, citation networks, diffusion (AIDS, news, innovations) networks, data-mining (2-mode networks), etc. The design of **Pajek** is based on development of graph data structure and algorithms libraries Graph and X-graph, collection of network analysis and visualization programs STRAN, RelCalc, Draw, Energ, and SGML-based graph description markup language NetML.

With **Pajek** one can: find clusters (components, neighbourhoods of 'important' vertices, cores, etc.) in a network, extract vertices that belong to the same clusters and show them separately, possibly with the parts of the context (detailed local view), shrink vertices in clusters and show relations among clusters (global view). Besides ordinary (directed, undirected, mixed) networks **Pajek** supports also multi-relational networks, 2-mode networks (bipartite (valued) graphs – networks between two disjoint sets of vertices), and temporal networks (dynamic graphs – networks changing over time).

One of its vital features which have been used in this work is determining the components of the complex network. For example, the various components found by Pajek in the integrated molecular network, includes.

1. Strong – Strong Components of selected network.

2. Weak – Weak Components of selected network.

It is also used to explore the bow-tie architecture of the integrated molecular network by finding the partition vertices of the directed network such as

1. Giant strong component
2. IN
3. OUT
4. TENDRILS
5. TUBES

The file type used by Pajek is NET and a sample net file looks like the following:

```
*Vertices      4
     1 "Andrej"   0.1201    0.2849    0.5000   ellipse
     2 "Vlado"    0.8188    0.2458    0.5000   box
     3 "Pajek"    0.3688    0.7792    0.5000   diamond
     4 "Book"     0.8359    0.8333    0.5000   triangle URL
*Edges
1 2
1 3
1 4
2 3
2 4
3 4
```

One of the important functionalities of Pajek includes the usage of macros which enables one to record a sequence of primitive Pajek commands into a file. You can use this file later to execute the saved sequence of commands without selecting one by one.

.

Ma, H. W., Kumar, B., Ditges, U., Gunzer, F., Buer, J. and Zeng, A. P. (2004c). "An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs." *Nucl. Acids Res.* **32**: 6643-6649.

Chapter 3

An Extended Transcriptional Regulatory Network of *E.coli*

3.1 Introduction

The study of genome-wide transcriptional regulatory network (TRN) has drawn much attention in the last few years because it offers the possibility to better understand the topology and function of gene regulation of cellular responses to environmental changes at a system level (Guelzim *et al.* 2002; Lee *et al.* 2002; Milo *et al.* 2002; Shen-Orr *et al.* 2002; Bar-Joseph *et al.* 2003; Gutierrez-Rios *et al.* 2003; Martinez-Antonio and Collado-Vides 2003; Babu *et al.* 2004; Kao *et al.* 2004; Luscombe *et al.* 2004; Herrgard and Palsson 2004b). A prerequisite for this kind of studies is the TRN reconstruction.

However, the reconstruction of genome-scale TRN is not an easy task. This is because: (i) we cannot directly obtain regulatory relationships from gene annotation information as in the case of the relationship between gene and metabolic enzyme (Ma and Zeng 2003a); (ii) computationally predicted relationships between transcription factors and their regulated genes by methods such as binding motif analysis are often not reliable due to the existence of short binding sequence that is not well conserved among organisms (Alkema *et al.* 2004; Yu *et al.* 2004; Herrgard *et al.* 2004a). A combination of genomic information with genome-wide expression data under diverse experimental conditions is often necessary and represents a promising approach (Ihmels *et al.* 2002; Gutierrez-Rios *et al.* 2003; Herrgard *et al.* 2003; Segal *et al.* 2003; Yu *et al.* 2003). Till now, the majority of studies of genome-scale TRN are focused on the two experimentally well-studied model microorganisms: *Saccharomyces cerevisiae* and Escherichia coli. Information on large-scale TRNs of other organisms is rather limited (Matys *et al.* 2003; Herrgard *et al.* 2004a). RegulonDB (Salgado *et al.* 2004) and Ecocyc (Karp *et al.* 2002a) are the two most prominent databases for *E.coli* regulatory network that collect information on regulatory relationships through literature study and manual curation to maintain a high-quality data. However, the reconstruction of the *E.coli* TRN from these databases is not straightforward because the gene–gene regulatory relationships are stored in different files. It is also difficult to compare the data in these two databases since the information is stored in different formats and a different gene nomenclature system is used. Based on RegulonDB and new information from literature, (Shen-Orr *et al.* 2002) compiled a list of transcriptional regulatory interactions in E.coli through which one can build the TRN (this network is termed ‘TRN-SO’ in the following parts). This dataset has been used in several recent structural studies of E.coli TRN (Shen-Orr *et al.* 2002; Dobrin *et al.* 2004; Ma *et al.* 2004b). These studies have revealed several interesting structural features of TRN. Network motifs, which are considered as the basic building blocks of TRN, have been identified and shown to have important specific functions and implications for the dynamic control of gene regulation (Milo *et al.* 2002; Shen-Orr *et al.* 2002; Wolf and Arkin 2003; Mangan and Alon 2003a; Dobrin *et al.* 2004). Dobrin *et al.* (Dobrin *et al.* 2004) further showed that the two previously identified motif types of TRN (i.e. the feedforward loop and the bi-fan motifs) aggregate into homologous motif clusters.

3.2 *E.coli* TRN

The databases Ecocyc (version 8.0, www.ecocyc.org) and RegulonDB (version 4.0, http://www.cifn.unam.mx/Computational_Genomics/regulondb/) are readily downloadable from the Internet. For the RegulonDB database, the gene–gene transcriptional regulatory relationships are extracted from six files: product_table.dat (relationships between gene and its coded polypeptide product), polyp_prot_link.dat (relationships between polypeptides and proteins), conformation_table.dat (the modified protein conformation), regulatory_interaction.dat (which promoter is regulated by which activated protein), transcription_unit.dat (which transcription unit does an operon belong to) and trans_gene_link.dat (which genes are in which transcription unit). Starting from one gene, all the corresponding genes regulated by are identified from the information stored in these files, thus making it possible to represent the TRN as a graph in which nodes represent genes and links as transcriptional regulations. Besides these files, the file 'promoter.dat' is also used to obtain more regulatory interactions for which the promoters do not belong to any known transcriptional units. In addition, except for sigma 70, all the other sigma factors are regarded as transcription factors and the interactions between them and their regulated genes are also added in the network. In this way, a network with 1024 genes and 2065 interactions is obtained. In contrast, the popularly used TRN-SO network contains only 855 genes and 1330 interactions between the genes (Shen-Orr *et al.* 2002).

For the database Ecocyc, the gene–gene regulatory relationships were extracted from three files, namely bindrxns.dat (protein–promoter and protein–binding site relationships), transunits.dat (which genes, promoters and binding sites are in a transcriptional unit) and proteins.dat (protein–gene information) respectively. It is found that there are many missing links in the database. For example, there are several promoters and binding sites that are regulated by proteins in the 'bindrxns.dat' file but not included in the 'transunits.dat' file. In this case, the genes corresponding to the promoters and binding sites directly from the files 'promoters.dat' and 'dnabindsites.dat' are obtained. Thus the resulted network from Ecocyc includes 959 genes and 2034 interactions among these genes. The interactions by the alternative sigma factors are also included in this network.

3.3 An Extended *E.coli* TRN

A major problem while comparing the TRN from different data sources is that they often use different gene IDs. RegulonDB assigns a new ECK number for each gene in the E.coli genome which is mainly based on the original genome annotation, the so-called 'Blattner ''b'' number system' (Blattner *et al.* 1997). However, this annotation has been extensively updated since the genome sequence is completed. A number of genes are deleted or merged and certain new genes are added. A more reliable and up-to-date annotation is the EcoGene database developed by Rudd (Rudd 2000). The EcoGene accession number (EG number) is used in Ecocyc for gene representation. However, there are also a lot of genes in Ecocyc that use a different nomenclature system. For example, many genes have IDs starting with G rather than EG. To avoid this confusion, all the genes in the regulatory network are mapped to genes in the EcoGene database. It is found that one gene (b0725, corresponding to ECK120000726 in RegulonDB and G6388 in Ecocyc) has no EG number because it has been removed in the new annotation. There are six pairs of genes that have replicate EG numbers (araH 1,2; gatR 1,2; gntU 1,2; ilvG 1,2; tdcG 1,2; phnE, b4103) because they are merged in the new annotation. By using the consistent gene ID, the three TRNs from different sources for the same organism *E.coli* K-12 are compared which revealed that common genes in the three networks (Ecocyc, Regulon DB and TRN-SO) make up only half of the total genes while the common interactions are only about one third of the total interactions. Therefore integrating information from different resources is important for obtaining a more complete TRN for *E.coli*. A combined network that includes all the 2624 interactions from the three data sets has been produced. In addition, this network is further extended by adding 23 additional genes and 100 regulatory relationships through literature survey. These new regulatory relationships are mainly involved in iron response and acid resistance (Dill *et al.* 2001; Shen-Orr *et al.* 2002). Specifically, a small regulatory RNA, ryhB, and eight genes regulated by it were added in the network (Dill *et al.* 2001). It is noticed that no small RNA regulated interactions are included in RegulonDB and Ecocyc though many small regulatory RNA have been identified in recent years (Pennock *et al.* 2002). One possible explanation is that the regulatory mechanisms for most regulatory RNAs are still not clear. Many of them may regulate at posttranscriptional level as antisense RNAs and

thus are different from transcription factors that regulate the target genes at transcriptional level. The extended TRN altogether includes 1278 genes and 2724 interactions (Fig. 7). Compared with the TRN-SO network, the new network contains one-and-a-half times more genes and more than twice the regulatory interactions. Although it is still not a complete network for the transcriptional regulation in *E.coli*, it provides more reliability for structural and functional analysis at genome level. Comparison of the structure of the new network with that of the previous one can help us to estimate to what extent the incompleteness of the information affects the results of structural analysis.

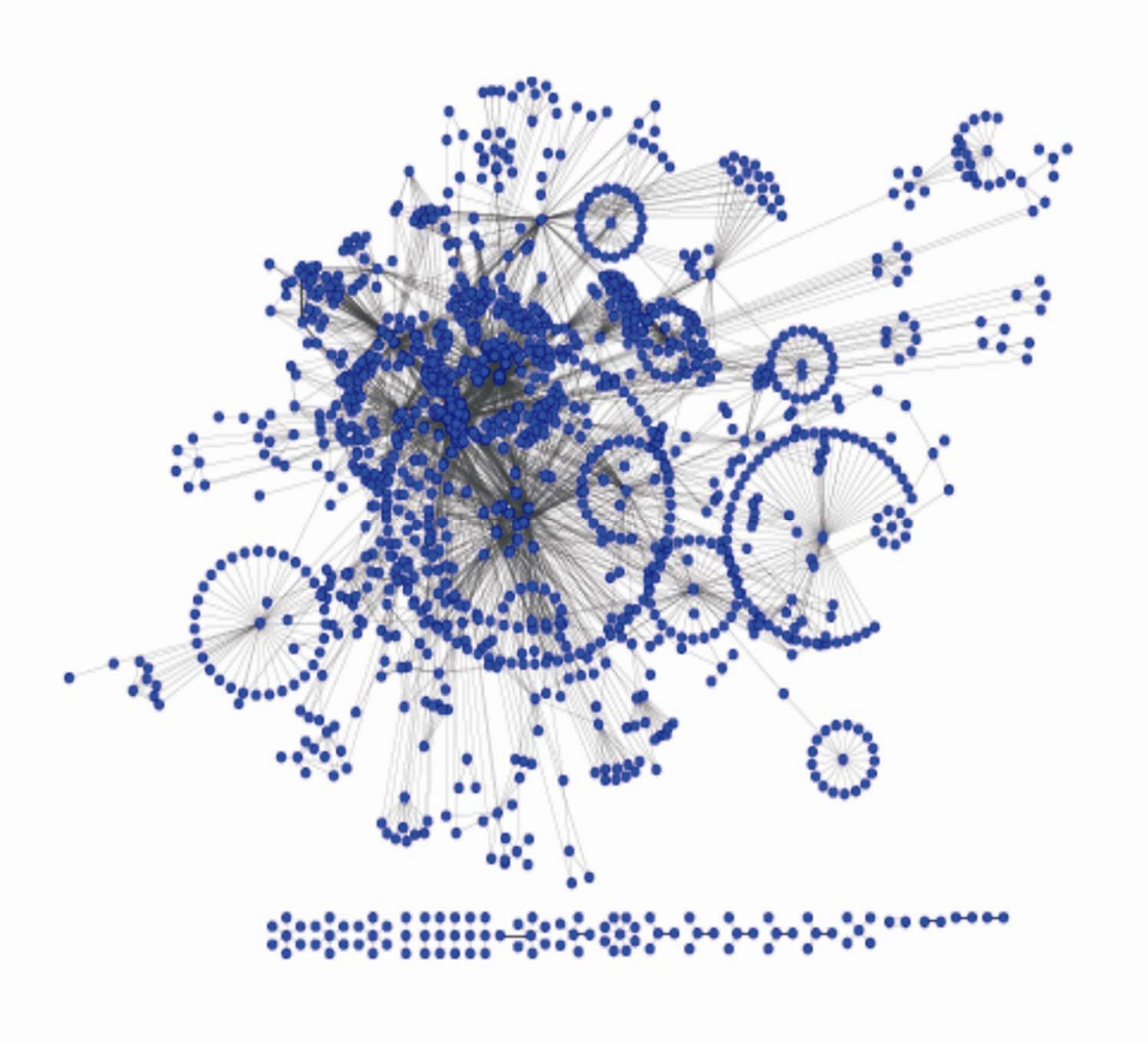


Figure 7. The extended transcriptional regulatory network of E.coli. The blue colored nodes represent genes and the grey colored edges represent transcriptional

regulation interactions (TRIs) network consists of 1278 genes (nodes) and 2724 TRIs (edges).

3.4 Multi-layer Hierarchical Structure of the Extended Network

To investigate whether the organizational structure of the combined network is in consistence with the previous TRN-SO network, the connectivity structure of the obtained network is analyzed using methods based on graph theory. Component analysis has revealed that there are seven two-gene regulatory loops (A regulates B and B also regulates A) in the new network, which is different from the previous finding based on TRN-SO (Ma *et al.* 2004b). It is found that both genes in the same loop are in the same operon and thus are regulated by the same set of transcription factors. This finding also explains why such loops aren't obtained during the previous studies while using operons as nodes (Ma *et al.* 2004b). Certain pairs of the genes in the loops code for different subunits of a transcriptional regulator such as *ihfAB*, whereas others may code for antagonistic regulators, which regulate almost the same set of genes. For example, *marA* codes for a transcription activator of the multiple antibiotic resistance locus and *marR* codes for a transcription repressor of the same genes. The two genes are in the same operon and both regulate the expression of this operon, resulting in a two gene regulatory loop. By placing the two genes in a loop at the same layer, a multi-layer hierarchical structure (Fig. 8) is obtained similar to that found previously (Ma *et al.* 2004b). However, nine layers instead of five are in the regulatory hierarchy. This result is not surprising if we consider the fact that the new network includes more interactions among the regulators than the previous one. Among the 14 regulators in the top five layers, six have been identified as global regulators in our previous studies (*crp, rpoS, ihf, cspA, hns and rpoN*). Four other regulators (*phoB, fis, soxR and rpoE*) have also been identified as global regulators in three previous papers (6, 59). DnaA is annotated as 'initiator protein for DNA synthesis and global transcription regulator' in Ecogene database (Rudd 2000). Therefore this result supports the conclusion that the top layer regulators tend to be global regulators. The confirmation of the multi-layer hierarchical structure in the extended TRN strongly implies that it is an underlying structure of the TRN in E.coli. A possible biological explanation for the existence of this hierarchical structure is that the

interactions in TRN are between proteins and genes. Only after a regulating gene has been transcribed, translated and eventually further modified by cofactors or other proteins, it can regulate the target gene. A feedback from the regulated gene at transcriptional level may delay the process for the target gene to access a desired expression level in a new environment. Feedback control may be mainly through other interactions (e.g. metabolite and protein interaction) at post-transcriptional level rather than through transcriptional interactions between proteins and genes (Wall *et al.* 2004; Ma *et al.* 2004b). For example, a gene at the bottom layer may code for a metabolic enzyme, the product of which can bind to a regulator which in turn regulates its expression. In this case, the feedback is through metabolite–protein interaction to change the activity of the transcription factor and then to affect the expression of the regulated gene. Therefore, to fully understand the gene expression regulation, an integrated network that includes different interactions is needed.

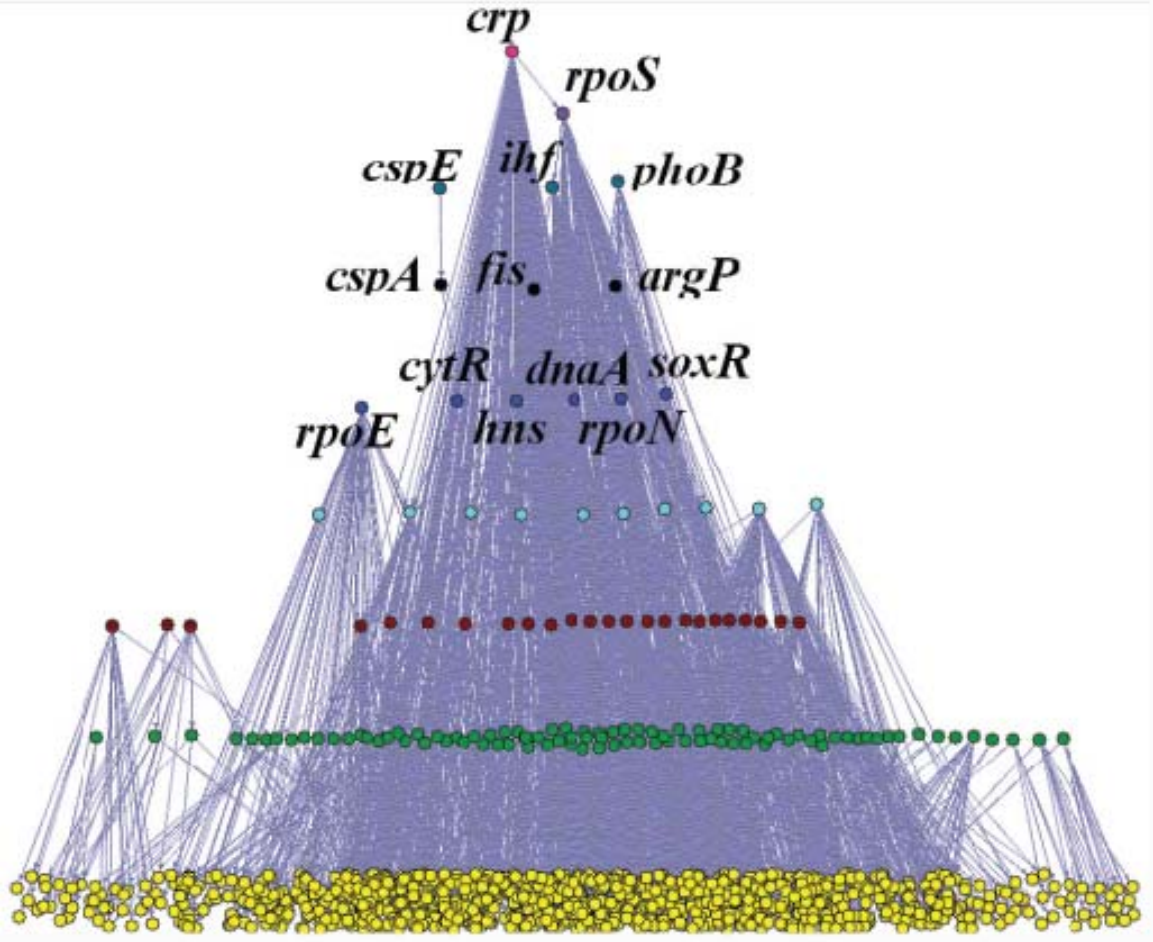


Figure 8. The multi-layer hierarchical structure of the extended TRN of E.coli.

3.5 Network Motifs and Motif Organization

To calculate network motifs in the E.coli TRN, all the loops in the network (including the auto-regulatory loops and the two-gene regulatory loops) are removed. The program Mfinder developed by Kashtan *et al.* (Kashtan *et al.* 2004) is then used to generate the motif profiles. In agreement with previous findings, feed-forward loop (FFL) is the only three-node motif (Ravasz *et al.* 2002). There are 712 FFLs in the network, far more than the 42 FFLs found in the TRN-SO network (Mangan and Alon 2003a). One reason for this large difference is the use of operons rather than genes as nodes in the TRN-SO network. When genes are used as nodes, 162 FFLs are obtained in the TRN-SO network. It is still less than one-fourth of that found in the new network. The first four types are the so-called coherent FFLs in which the direct effect of the up regulator is consistent with its

indirect effect through the mid regulator (Mangan and Alon 2003a). In contrast, the last four types of FFLs are incoherent because the direct effect of the up regulator is contradictive with its indirect effect. The total number of incoherent FFLs is 152, which is only a little less than half of the number of the coherent FFLs (330). This result is inconsistent with the result from the TRN-SO network where only 14 of the 162 FFLs are incoherent (7 of the 42 FFLs are incoherent while using operons as nodes), but very similar to the network of *S.cerevisiae* (25 of 56 FFLs are incoherent) studied by the same authors (Bateman *et al.* 2004). Another interesting point is that the first and the fifth FFL types predominate the coherent and incoherent FFLs respectively in both networks of *E.coli* and *S.cerevisiae* (Mangan and Alon 2003a). These results indicate that the distribution and predominance of FFLs in the TRNs of both E.coli and *S.cerevisiae* have similar patterns. Further examination revealed that the distribution of different types of FFLs for various regulators. Most of the regulators are found to regulate only one or two types of FFLs. For example, *flhDC, lysR, soxS, rob* and *tdcR* mainly regulate type one FFL, while *modE* regulates type five FFL and *cpxR* regulates type two FFL. Most of the type four FFLs and the type eight FFLs are regulated by fnr. Mangan *et al.* (Mangan and Alon 2003a; Mangan *et al.* 2003b) studied the dynamic behavior of these different types of FFLs. They found that the incoherent FFLs could speed up the responses of the target gene while the coherent FFLs delay the response. The feature of less feedback regulation at transcription level as demonstrated by the multi-layer structure may imply another possible function of incoherent FFLs. By activating a gene and at the same time activating a regulator which represses the target gene, the upper regulators can control the gene expression at a proper level. More in-depth studies may help to examine whether such a mechanism is a general mechanism for gene expression regulation in TRN.

In a recent study (Dobrin *et al.* 2004), it is showed that network motifs are organized in a hierarchical way in the E.coli regulatory network: first, interacting motifs form motif clusters; motif clusters of different motifs are then connected to make a motif super cluster which is regarded as the backbone of the whole network. However, this concept of network motif organization is not valid for the extended network. Of the 712 FFLs, 701 are connected to form a giant motif cluster with 435 genes, while the remaining 11 FFLs form four very small clusters. The reason for this large discrepancy is that the new

extended network includes more interactions and thus obtains more motifs that can link the previously disconnected motif clusters together. Therefore, caution should be taken in dealing with results from an incomplete network, especially when drawing conclusions about general organizational principle(s).

3.6 Genes Regulated by Interacting Network Motifs

As mentioned above, FFLs are considered to have important functions in controlling the dynamic response of the target gene (Mangan and Alon 2003a). Therefore it is of interest to check how many genes in the TRN of *E.coli* are regulated by FFLs and by how many FFLs. It is found that there are altogether 400 genes that are regulated by one or more FFLs in the network. Among them, 383 genes are at the bottom layer of the hierarchical structure; they account for only about one-third of the total genes (1121) at the bottom layer. Furthermore, only 56 genes are solely regulated by one FFL. All the other genes are regulated by two or more FFLs or by one FFL together with certain other regulators that do not form a FFL. There are also a few genes (such as *cysG*, *glpAB* and *nirB*) that are regulated by both coherent (delaying the response) and incoherent FFLs (speeding up the response). These results indicate that the previous studies on dynamic behavior of FFLs may be pertinent for only a small part of the genes in the network. It would be interesting to examine the dynamics of target genes that are controlled by several interacting motifs and also by other regulators. Figure 9A and B depict two examples of complex regulatory circuits in which the target gene is regulated by six and five FFLs respectively. In Figure 9A, the target gene gadA codes for glutamate decarboxylase, an important metabolic enzyme in the gammaaminobutyric acid (GABA) shunt which is important in oxidative stress response of plant cells (Bouche and Fromm 2004) and bacteria and is also an important component in the acid resistance system of *E.coli* (Shen-Orr *et al.* 2002). In Figure 3B, the target gene *lpd* codes for lipoamide dehydrogenase which is a component of the pyruvate and 2-oxoglutaratedehydrogenase complexes. Both the pyruvate and 2-oxoglutarate dehydrogenase complexes play key roles in the metabolism of *E.coli*. lpd also functions as glycine cleavage system L protein. The multiple and important functions of these target genes may explain why they are controlled by several interacting FFLs. In these complex regulatory circuits, the dynamic

expression pattern of the target gene will be obviously different from that controlled by one FFL.

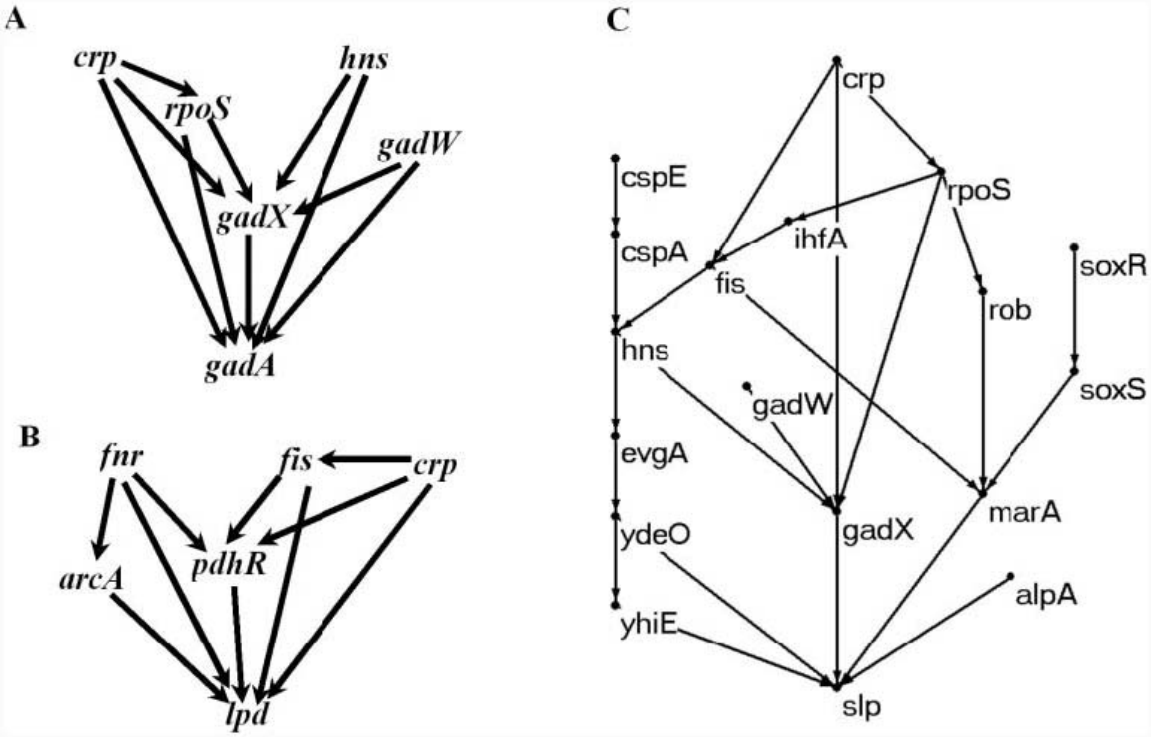


Figure 9. Example of complex regulatory circuits.

(A) Gene *gadA* is regulated by six FFLs; (B) gene *lpd* is regulated by five FFLs; and (C) gene *slp* is regulated by 17 regulators.

Furthermore, the target gene may also be controlled by regulators not belonging to any FFLs. A total of 692 genes are regulated by more than two regulators. The most complex regulatory circuit is the one for the gene *slp*, which codes for an outer membrane lipoprotein induced under carbon starvation and stationary phase (Rudd 2000). It is regulated by 17 regulators as shown in Figure 9C. These regulators participate in cellular responses to various environmental conditions, such as oxidative stress (*soxRS*), acid stress (*gadW, gadX, evgA, ydeO* and *yhiE*), cold shock (*cspE, cspA*) and multiple antibiotic resistance (*marA*). This underlies the importance of this gene in stress response. Further studies are required to elucidate the exact function of this gene. Although *slp* is regulated by only one FFL (*ydeO-yhiE-slp*), it is in fact also regulated by other FFLs that contain more than three genes. For example, *fis-hns-evgA-ydeO-marA-slp* forms a six

node FFL, while *rpoS-rob-marA-gadX-slp* makes a five node FFL. One common feature of these FFLs (including the three node FFL) is that one top regulator regulates a target gene through different pathways. Further studies are required to understand the dynamic response of genes under the control of different loops and different regulators. More care should be taken when studying gene expression dynamics in these complex regulatory circuits. It may be completely different from that of a simple three node FFL. Furthermore, only a subset of the regulators (and their regulatory interactions) in the whole network is activated under given environmental conditions as recently shown by Luscombe *et al*. (Luscombe *et al.* 2004). Therefore, for the complex regulatory circuits shown in Figure 3, it is possible that only one or two motifs are active at a given time. Further studies are required to investigate the dynamic change of the network topology and its effects on gene expression dynamics. Network motif analysis has provided useful information about the gene regulation patterns. It would be interesting to know the dynamic feature and the overall effect of individual motifs if they interact with other motifs within a more complex network and in a dynamically changed environment. The complex regulatory circuits as shown in Figure 9 may be good examples for such studies. Recently, Van Nimwegan (Van Nimwegen 2003) found by comparative genome analysis that the number of genes in each functional the genome. Specifically, the exponent for transcriptional regulators is almost two, implying that the number of transcription factors increases much faster than the size of the genome (quadruple transcription factors in a doubled genome).

This finding was further verified by Ranea *et al*. (Ranea *et al.* 2004) who investigated the distribution of protein super families in 56 different bacterial species. The non-linear scaling observed by these authors could be explained by an increase in complex inter-regulation of transcription factors as illustrated in the examples in Figure 3 and by the multi-layer hierarchical structure shown in Figure 8. The inter-regulation among the transcription factors can lead to a faster growing number of regulators than the number of genes with increase of the genome size.

Chapter 4

Integrated Molecular Network of *E.coli*

4.1 Introduction

Cellular function is normally coordinated by interactions between different cellular components and at different molecular levels. Therefore, to gain a more comprehensive picture of various cellular processes and their regulation, an integrative approach is required. A promising solution is to integrate the molecular networks at different molecular levels and to study them from a systems point of view. Recent development in high-throughput studies of genome-wide transcriptional regulatory network, protein-protein interaction (PPI) network and metabolic network (Kanehisa and Goto 2000; Karp *et al.* 2002a, 2002b, 2005; Rudd 2005) makes this possible.

The necessity of an integrated approach can be well exemplified by the widely studied model organism *E. coli*. In two of our recent studies, we investigated the organization and functionality of the genome-scale metabolic network of *E. coli* (Ma *et al*. 2003a, b). In another study (Ma *et al*. 2004c), we extended the transcriptional

regulatory (TRN) network of *E. coli* and analyzed its global structure and network motifs. An acyclic multi layer hierarchical structure of the TRN was discovered. This acyclic structure can be attributed to the fact that there are no feedback regulations at the transcriptional level. This was unexpected in view of the fact that feedback regulation is one of the fundamental regulation mechanisms of biological processes (Schuster *et al*. 2000). On the other hand, it is understood that the TRN is connected to the metabolic network and other types of molecular networks which may mediate the feedback regulation of gene expression. In order to uncover the complexity of a biological system and thereby identify mechanisms such as feedback regulation of gene expression, it is important to integrate transcriptional regulation interactions with metabolic reactions and metabolite-protein (specifically metabolite-transcription factor) interactions.

4.2 Methods for Network Integration and its Structural Analysis

4.2.1 Data Sources

The integrated molecular network (IMN) of *E. coli* was assembled with three types of interactions: transcriptional regulation interaction, metabolic reaction and metabolite protein interaction. To this end, we extracted the transcriptional regulation data available from the recent study (Ma *et al*. 2004c), where an extended transcriptional regulatory network of *E. coli* was reconstructed using the databases RegulonDB (http://cifn.unam.mx/Computational_Genomics/regulondb/) (Salgado *et al*. 2004), Ecocyc (http://www.ecocyc.org) (Karp *et al*. 2002a), the data provided by Shen-Orr *et al*., and additional literature data. The network includes 1278 genes and 2724 interactions. For metabolic reactions, we used the *E. coli* network reconstructed by Ma and Zeng (Ma and Zeng 2003a). A large number of genes encode for enzymes which catalyze metabolic reactions and some of the metabolite products act as co-factors interacting with transcription factors, thereby acting as an inducer or a co-repressor. We extracted these interactions between metabolite and proteins (transcription factors) from the databases Regulon DB, Ecocyc and EcoTFs (http://ecotfs.lanl.gov/table1.html) (Wall *et al*. 2004).

4.2.2 Graph Representation of the IMN

The resulting integrated molecular network of *E. coli* consists of 1549 genes connected by 6497 interactions of three types: (i) transcriptional regulation interaction (TRI); (ii) metabolic reaction (MR); and (iii) metabolite protein interaction (MPI) (Fig. 10). All the nodes in the network represent genes whereas the edges represent various interactions. In order to differentiate the one interaction from the other since they are completely different from each other, different colors are used. To illustrate in detail, when the product of a gene *x* regulates the expression of gene *y*, then gene *x* (node A) is connected to gene *y* (node B) and the transcriptional regulation interaction between these two genes is represented by a grey colored edge.

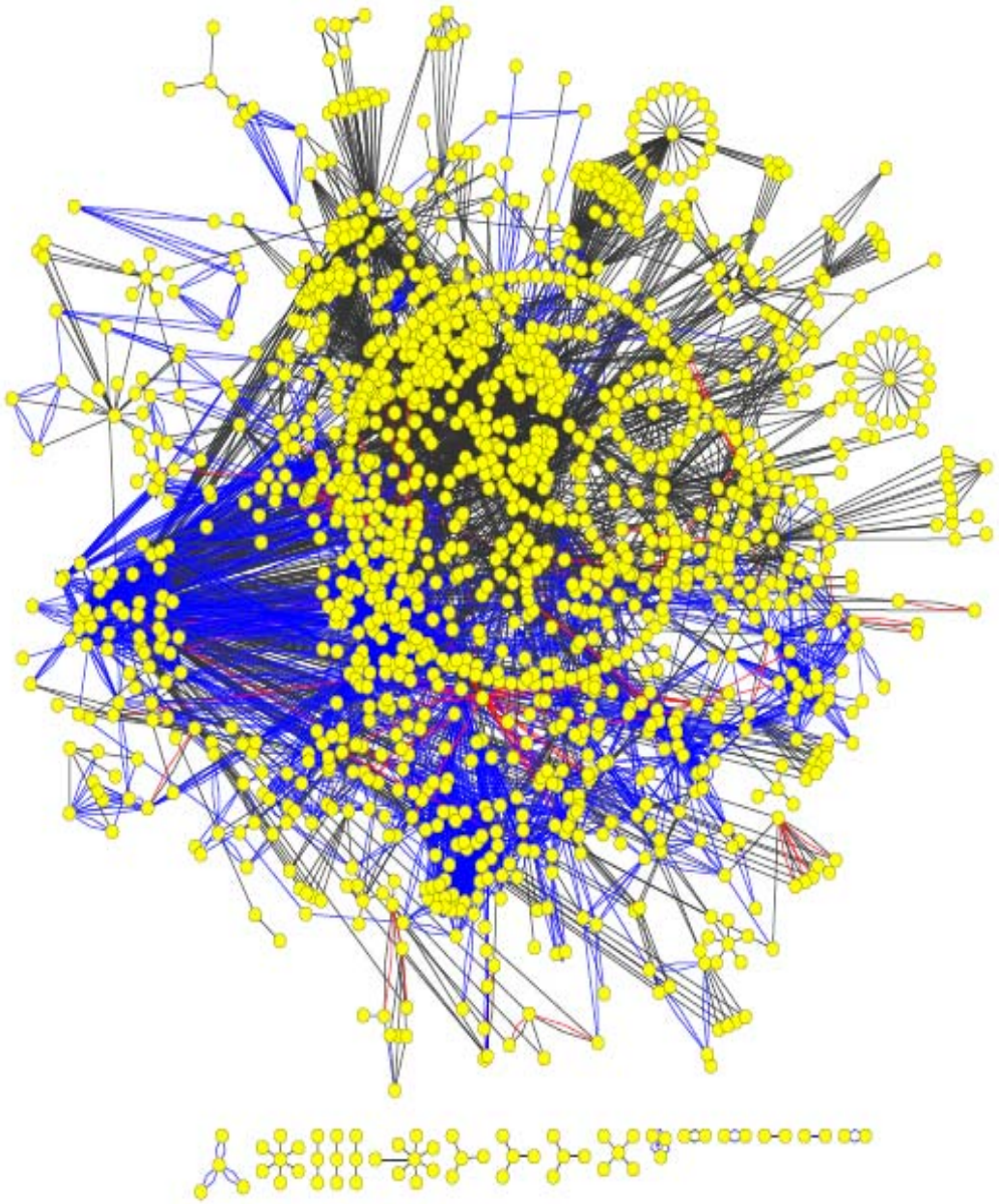


Figure 10. The integrated molecular network of E. coli.
The nodes represent genes, the grey edges represent transcriptional regulation interactions; the blue edges represent the metabolic reactions and the red edges represent the metabolite-protein interaction.

interaction when the gene *x* encodes for a catabolic or biosynthetic enzyme catalyzing a metabolic reaction with metabolite m as a product that acts as a co-factor for a

transcription factor z. The various interactions in the network are represented graphically as directed links of three colors: black links for TRI, blue links for MR and red links for MPI. The software Cytoscape (http://www.Cytoscape.org/) (Shannon *et al*. 2003) was used for the network representation.

4.2.3 Detecting Feedback Loops

The main idea of integrating metabolite-protein interaction with other interaction types like transcriptional regulation and metabolic reaction is due to the conclusion made by Ma *et al*. 2004c, where it was suggested that feedback in the TRN may be mediated through MPIs from the genes at the lowermost layer to those in the upper layer of the hierarchical structure. In order to find out these feedback regulations first we searched for the entire genes at the lower layer in the hierarchical structure of the TRN which code for metabolic enzymes. Then we looked for the MRs catalyzed by these enzymes and retrieved a list of products produced by them. Finally we searched for the products that bind to regulators which regulate the expression of genes at the upper hierarchical layer that are responsible for regulating the metabolic genes at the lowermost layer of the extended TRN. For example, the transcription factor TreR regulates two genes *treB* and *treC* in the trehalose degradation pathway. One of the regulated genes, *treB*, codes for a TreB-monomer which forms the protein complex EIITre. EIITre catalyzes the transport reaction by the PTS transport system (Fig. 11):

Phospho-enol-pyruvate + Trehalose = Trehalose-6-phosphate + Pyruvate

One of the metabolic products of the reaction, trehalose-6-phosphate is a cofactor for TreR. Therefore through trehalose-6-phosphate a two-node FBL between *treB* at the lowermost layer and *treR* at the seventh layer of the hierarchical structure of the extended TRN (Ma *et al*. 2004c) is formed.

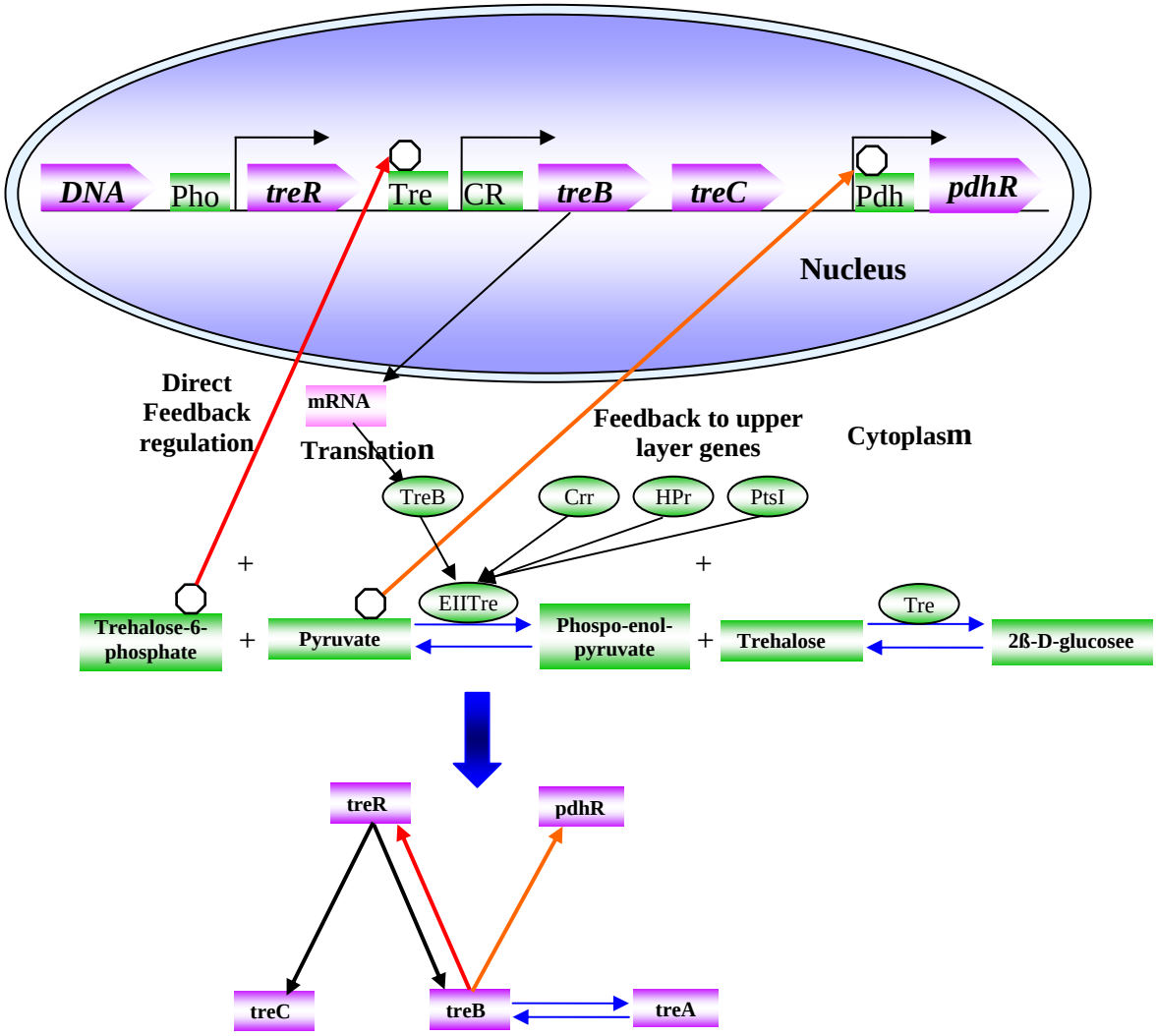


Figure 11. Schematic representation of the IMN of E.coli involving multiple interaction types.
(a) shows transcriptional regulation of proteins at the gene level, the proteins involved in catalyzing the metabolic reaction where ß-D-glucose is converted to trehalose which eventually is converted to pyruvate and trehalose-6-phosphate and the binding of these metabolites to the proteins regulating the feedback. (b) Simplified representation of IMN. The gene *treR* regulating the transcription of *treB* and *treC* is represented as black edges. The genes *treR* and *treA* are linked using blue edges since they produce the metabolic enzymes EIITre and Tre which catalyze two successive reactions having common metabolites (phosphor-enol-pyruvate and trehalose). The metabolic products pyruvate and trehalose-6-phosphate act as cofactors for the transcription factor TreR and PdhR and these metabolite protein interactions are represented as red edge (direct feedback from *treB* to *treR*) and orange edge (from *treB* to *pdhR*).

4.2.4 Identification of Three-gene Network Motifs

We analyzed the IMN for three-node network motifs (Milo *et al*. 2002; Shenn-Orr *et al*. 2002) using the method of Yeger-Lotem *et al.* 2004. The algorithm executes the multi-edged IMN by generating 1000 randomized networks and detects patterns which recur significantly much often in the real network than in the randomized networks. The minimum number of occurrences is considered to be 10 or above. For statistical significance only those patterns with a *P*-value less than 0.001 are considered. Here the *P*-value is defined as the ratio between the number of randomized networks in which the pattern appears as often as in the IMN to the total number of random network generated.

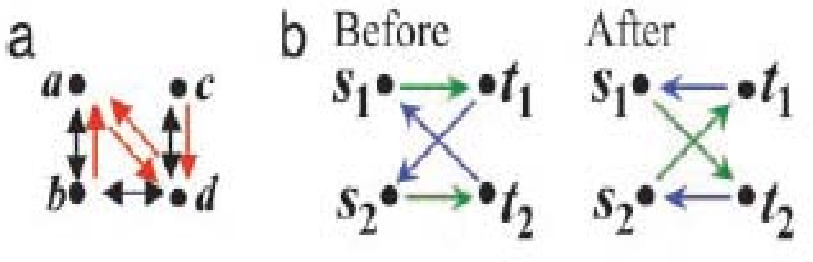


Figure 12. The randomization procedure.
(*a*) Extended node degree and edge profile. Nodes represent proteins; black, bidirected edges represent PPIs; and red, directed edges represent TRIs. Extended node degrees: *a*, one PPI, one outgoing TRI, and two ingoing TRIs; *b*, two PPIs and one outgoing TRI; *c*, one PPI and one outgoing TRI; *d*, two PPIs, one outgoing TRI, and two ingoing TRIs. Examples for edge profiles: (*a*, *b*), one PPI and one ingoing TRI; (*b*, *a*), one PPI and one outgoing TRI; (*a*, *d*), one outgoing TRI and one ingoing TRI; (*b*, *d*), one PPI. The edge profile of (*d*, *c*) is equivalent to that of (*a*,*b*). (*b*) The four-point-switchability condition. If edge profile (*s*1, *t*1) = edge profile (*s*2, *t*2) and edge profile (*s*1, *t*2) = edge profile (*s*2, *t*1), then edges can be switched as exemplified. For clarity, each edge color represents a type of edge profile. Note that if (*s*1, *t*1; *s*2, *t*2) are switchable, then so are (*s*1, *t*2; *s*2, *t*1), (*t*1, *s*1; *t*2, *s*2), and (*t*2, *s*1; *t*1, *s*2). Switchability is considered only for cases in which all four nodes are distinct and at least one edge profile is not empty.

To generate randomized networks with multiple types of interaction, two terms are considered by this algorithm:

1. Extended degree of a node: the number of edges per type that point to from a node (demonstrated in Fig. 12*a*). Two nodes have the same extended degree if they have the same number of ingoing and outgoing edges for each edge type.

2. Edge profile of two nodes: the set of edges connecting the two nodes with the type and direction of each edge detailed. The local connectivity of a node is reflected by the extended degree whereas edge profile provides a local measure of the relation between the nodes. The randomized networks are generated such that both the extended degree of each node and the profile of each edge in the network are retained. By an iterative switching of edges, this algorithm generates such random networks. Sufficient conditions for the retention of all edge profiles and the extended degrees of all nodes are provided by the four-point-switchability condition. The network motifs detected consisted any of the three interactions: TR, MR and MPI or a mixture of them. We then searched for the genes forming the motifs of the IMN which generated a list of distribution of motifs within the GSC of the bow-tie structure of the IMN.

4.2.5 Discovery of the Bow-tie Structure

The IMN is comprehensively analyzed to discover the existence of bow-tie structure with a giant strong component (GSC). We found that there exist several fully connected sub-networks in the IMN and within each of these subnetworks; one node can be reached by another node within the same subnetwork. These fully connected sub-networks are called strong components and the largest of them is called giant strong component through which connections to other components of the bow-tie architecture is established. In graph theory, a strong component of a network is defined as a subset of nodes such that for any pair of nodes *u* and *v* in the subset there is a path from *u* to *v*. Other components found in the bow-tie structure are IN, OUT, Tubes and Tendrils (Fig 13). The IN component comprises nodes which connect the nodes in the GSC and the direction of the edge connecting these two components is always from IN to GSC. OUT subset contains nodes which are connected to GSC and the edge direction is always from GSC to OUT.

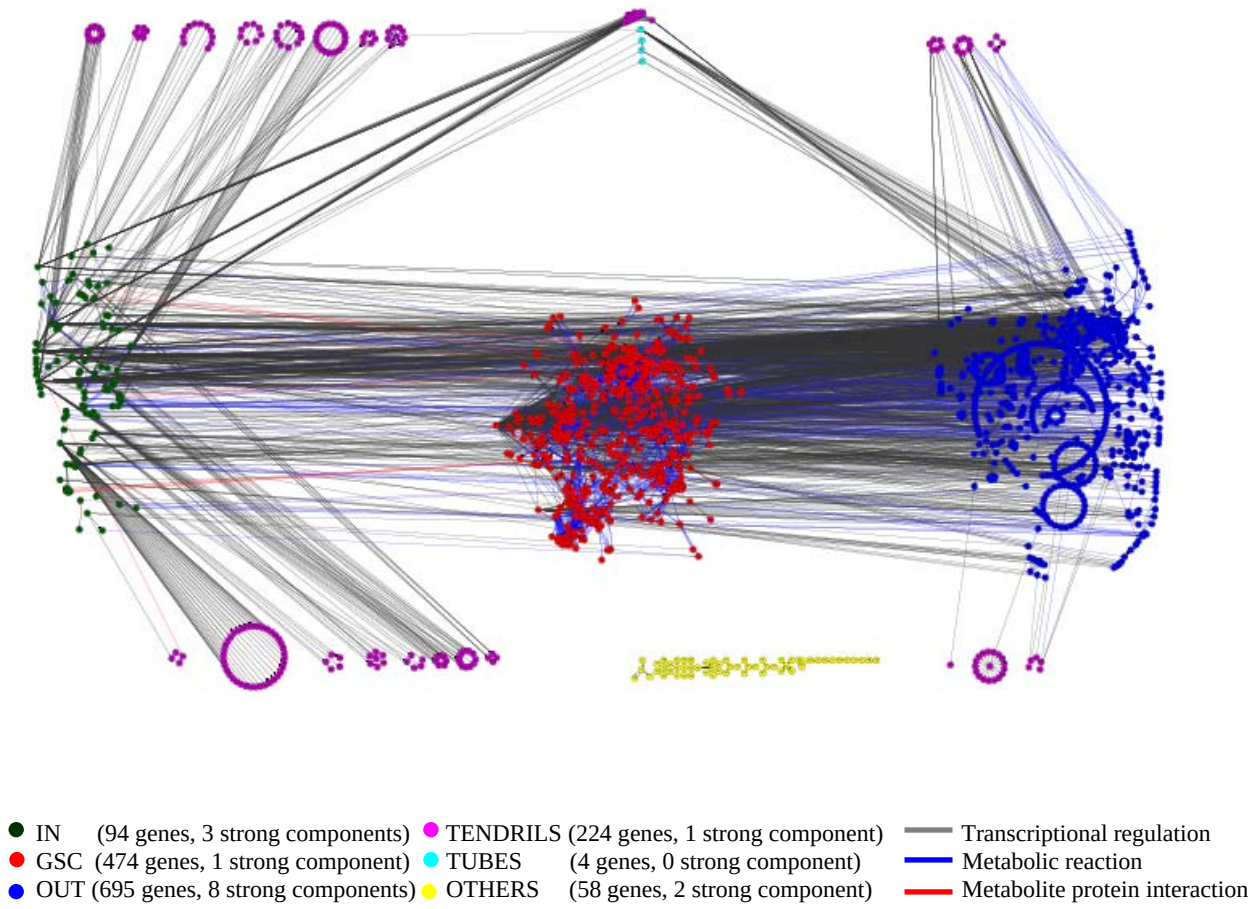


Figure 13. The bow-tie connectivity structure of the IMN of E. coli.
The nodes represent genes and the green nodes represent the IN subnetwork (nodes with directed link to reach the connected core, but cannot be reached from it); the red nodes represent the giant strong component (GSC – the knot of the bow-tie which contains several highly connected nodes such that each node can be reached from the other through one or several paths); the blue nodes represent OUT subnetwork (which contains nodes that are linked from nodes in GSC, but not linked back to it); the pink nodes represent TENDRILS (nodes that are linked from IN or linked to OUT subnetworks, but are not directly connected to the GSC); the yellow nodes represent OTHERS subnetwork (disconnected part of the network) and the turquoise nodes represent TUBES (connects the IN and OUT subnetworks and not connected to the GSC). The grey edges represent transcriptional regulation interactions; the blue edges represent the metabolic reactions and the red edges represent the metabolite-protein interaction.

TUBES connect genes from IN to OUT and there is no linkage through GSC. TENDRILS contain nodes which are loosely connected to IN and OUT components. All

the components in the integrated molecular network of *E.coli* were identified by using the network analysis software Pajek (Batagelj and Mrvar 1998).

4.3 Feedback Regulations

In our previous study of the regulatory network of *E. coli* (Ma *et al*. 2004c), it was shown that there are no FBLs at the transcriptional level as revealed by an acyclic multilayer hierarchical structure of the network. With the integrated network, we found that feedback regulations in *E. coli* are primarily mediated by metabolic-protein interactions where a gene at a lower layer in the hierarchical structure of the TRN code for a metabolic enzyme, the product of which can bind to a regulator which in turn regulates its expression (Fig. 14).

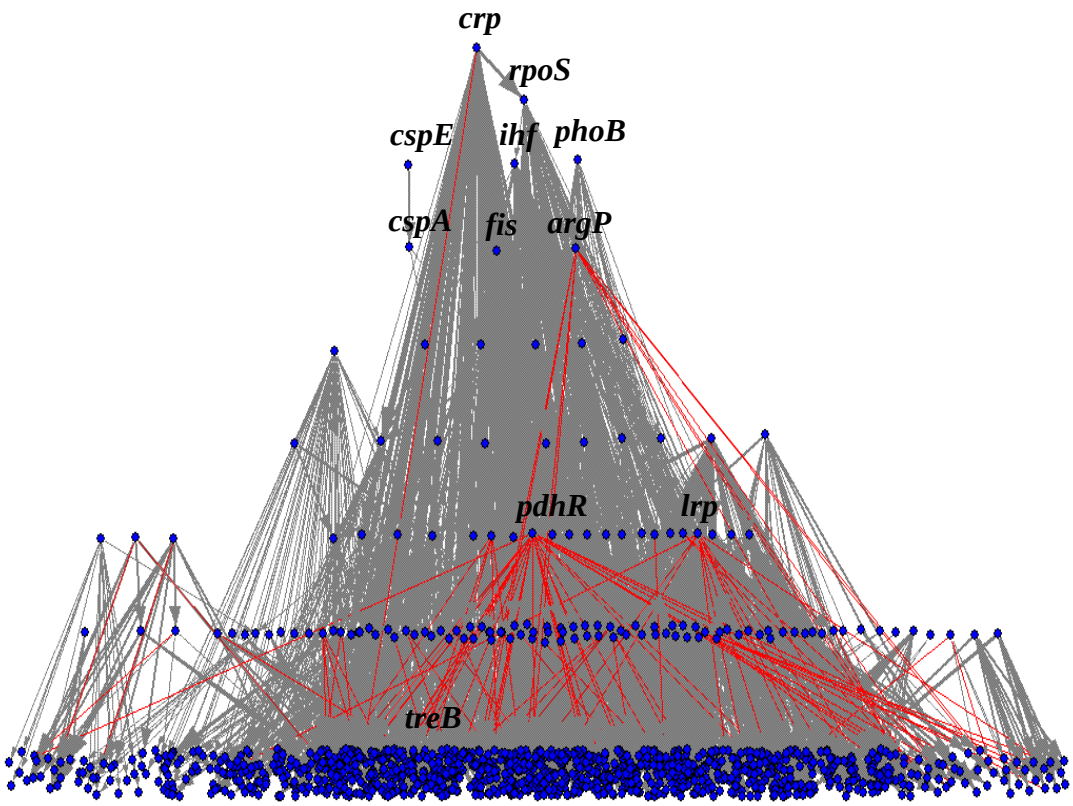


Figure 14. The multi-layer hierarchical structure representing the extended E. coli TRN(grey edges) and the feedback regulations through metabolite protein interactions (red edges).

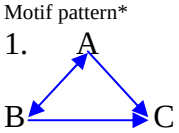
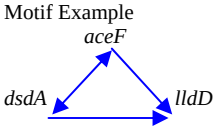
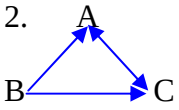
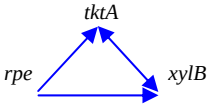
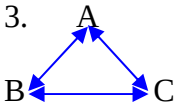
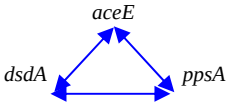
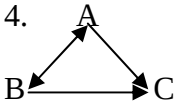
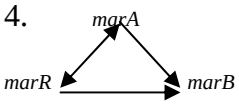
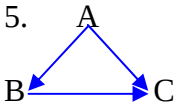
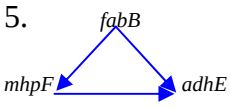
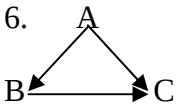
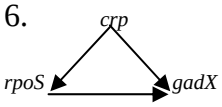
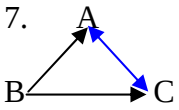
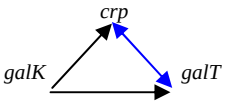
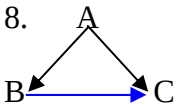
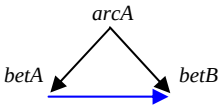
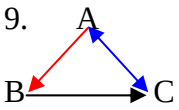
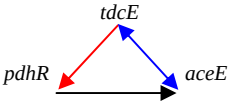
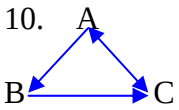
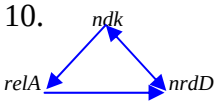
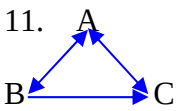
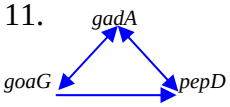
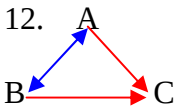
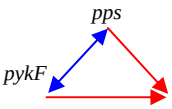
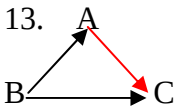
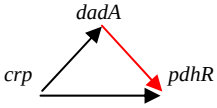
Here we define such regulations as feedback loops (FBLs). The feedback on transcriptional regulation comprises an enzyme catalyzed metabolic conversion and a metabolite-protein interaction excluding those involving currency metabolites such as ATP or NADH (see Methods). This is because the use of currency metabolites to form a possible feedback regulation of the regulator gene is not specific and the feedback regulation so inferred is biologically not meaningful in analogy to the use of currency metabolites as connectors for topological analysis of metabolic networks (Ma and Zeng 2003b). Systematically we identified 33 two-node FBLs and 36 three-node FBLs in the IMN and most of these FBLs are found to occur in several metabolic pathways like glycolysis, amino acid and arginine biosynthesis respectively. Majority of them are found to occur with the genes *pdhR*, *lrp* and *argP* and one of the reasons is that the metabolites which influence how these TFs affect transcription are produced through many reactions catalyzed by various enzymes at the lower layer of the hierarchy. It is remarkable that 24 genes present in the top six layers which include 11 global regulators (Chuang *et al*. 1993; Ma *et al*. 2004c; Martinez-Antonio and Collado-Vides 2003) exhibit only a few feedback regulations whereas the genes *crp* and *argP* are the only global regulators that have feedback regulations. The results presented above demonstrate that integration of different interactions can help systematically identify feedback regulations at the post-transcriptional level and their role in the molecular network.

4.4 Structural Analysis

4.4.1 Three-gene Network Motifs

The approach of Yeger-Lotem *et al.* 2004 was applied to analyze network motifs (Milo *et al*. 2002; Shenn-Orr *et al*. 2002) comprising multiple types of interactions (Yeger_lotem *et al*. 2004; Zhang *et al*. 2005) in an integrated network. Altogether 55 three-gene subgraphs were found in the network and of these 13 statistically significant ($P< 0.001$) subgraphs were identified as network motifs. The thirteen network motifs consist of eight patterns comprising either MR or TRI, four comprising a combination of two types of interactions among TRI, MR and MPI and only one pattern occurring with all three types of interactions. Table 1 provides a detailed list of all the network motifs that were detected in the IMN.

Table 1. Three-gene network motifs in the IMN of E. coli

Motif pattern*	Motif Example	Zscore†	Nreal	RndAvg	RndStd	Ngsc‡	Fraction§
1. 		201	2256	94	10.8	2032	0.90
2. 		186.1	625	10	3.3	581	0.93
3. 		184.5	3982	259	19.6	3901	0.98
4. 	4. 	29.5	199	5.2	6.6	37	0.18
5. 	5. 	26.3	344	89	9.7	289	0.84
6. 	6. 	16.1	818	257	34.9	176	0.21
7. 		15.7	280	86	12.3	166	0.59
8. 		14	221	74	10.5	130	0.58
9. 		13.8	21	1.8	1.4	21	1.00
10. 	10. 	11.2	100	31	6.1	100	1.00
11. 	11. 	10.1	272	131	13.9	272	1.00
12. 		5.2	13	3.2	1.9	13	1.00
13. 		3.8	24	8.6	4	22	0.91

*The edge: red represents MPI, blue represents MR and black represents TR respectively.

†Z-score= (Nreal - RndAvg)/ RndStd.

‡Number of motifs occurring in the GSC of real network.

§Ratio of Ngsc to Nreal.

We analyzed the distribution of all these motifs in the bow-tie structure of the IMN and found that about 80% of them appeared within the GSC. The motifs involving mixed type of interactions (Motif 7, 8, 9, 12 and 13) are mostly found in the GSC (Table 2). The motif patterns 1, 2, 3, 5, 10 and 11 are all composed of metabolic reactions. Of these one highly abundant motif is a metabolic clique (Table 1, motif pattern 3). This motif contains three metabolic enzymes and each of them catalyzes one of the three reversible metabolic reactions producing or utilizing a common metabolite through which all the other metabolites could be reached. For example, the products of the metabolic genes *dsdA, eda* and *malY* catalyze the reactions D-serine deaminase reaction, oxaloacetate decarboxylase reaction and cystathionine-beta-lyase reaction respectively. The metabolic products of these reactions are pyruvate, D-serine, oxaloacetate and cystathionine and each of these metabolites can be produced from each others through the common metabolite pyruvate. The metabolic clique motif occurs 3982 times in the network which accounts for nearly 44% of all the highly significant motifs in the network. The reason for this high abundance is due to the presence of hub metabolites which dominate the network via their highest degree of connectivity. In the above example, pyruvate which is produced through several reactions acts as a hub metabolite.

The motif pattern four represents two genes regulating each other forming a loop that co-regulates a target gene. The coregulating gene pair appears in two forms: (i) one of the coregulating genes acts as transcriptional activator and other a repressor (ii) both the genes are transcriptional activators and their transcription factors interact physically to regulate the target gene. Six sets of coregulating genes are found and the genes belonging to four of these sets (*glnG-glnL, rhaR-rhaS, gutM-srlR* and *marA-marR*) are found in the same operon. As an example of this type of motif, *marA* and *marR* belonging to the same operon marRAB regulate each other and these genes co-regulate the transcription of the target gene *marB.*

The motif pattern 6 describes the transcriptional feed-forward loop (FFL) which has been previously well studied (Mangan and Alon 2003a; mangan *et* al. 2003b; Milo *et al*. 2002; Shenn-Orr *et al*. 2002). It comprises transcription interactions between three genes where one gene regulates the other and these two genes coregulate the transcription of the target gene. For example, *crp* regulates the transcription of *rpoS* and both of these

global regulators together regulate several genes involved in several pathways.

The motif patterns 7 and 8 involve TRIs and MRs which consist of a pair of coregulated genes whose protein products (enzymes) catalyze successive metabolic reactions. In case of the motif pattern eight, the protein products catalyze irreversible metabolic reactions. In many of the cases the products of the coregulated genes form different protein complexes which catalyze the successive enzyme reactions. The ideal example for motif seven is *crp-galK-galT*: in this motif, the pair of genes *galK-galT* is coregulated by the global regulator *CRP* and the coregulated genes synthesize enzymes which catalyze successive reversible metabolic reactions in the galactose degradation pathway, whereas the best example for motif eight is *arcA-betA-betB* which is found in the betaine biosynthesis. The fact that coexpressed genes are coregulated (Yeger-Lotem *et al*. 2004) suggests that the metabolic genes catalyzing the reactions in a specific pathway should be coregulated at the transcriptional level. Further analysis of co-regulation with respect to metabolic pathways should be of great importance in understanding the complex regulatory circuitry of metabolism.

The motif pattern 9 is a FBL which comprises all the three interactions: TRI, MR and MPI. The function of this motif is well explained by an example *pdhR-aceE-tdcE*: In this motif, the protein PdhR regulates the transcription of *aceE* and then the metabolic proteins AceE and TdcE individually catalyze two successive reactions which bridge the TCA cycle, glycolysis and anaerobic glucose fermentation pathways. The final product ‘pyruvate’ acts as a cofactor for the transcription factor PdhR thereby mediating a FBL through metabolite protein interactions. Even though this motif is highly significant, its occurrence is very limited in the real network (<1%).

The motif pattern 12 is composed of MRs and MPIs which contain two genes whose protein products catalyze metabolic reactions producing the same metabolites which act as cofactors for the protein product of the regulatory gene. One such example comprises the metabolic genes *deoD and hpt*, the products of which catalyze the enzymatic reactions: inophosphor reaction and hypoxanthine phosphoribosyltransferase reaction respectively producing the product hypoxanthine which acts as a cofactor for the transcriptional factor *PurR.*

The motif pattern 13 contains two TRIs and one MPI which involves a pair of

genes that is coregulated by a common transcription factor. The protein product of one of the coregulated gene pair is involved in metabolite protein interaction with the protein product of the other gene. For example, *crp* co-regulates several gene pairs and one such gene pair includes *dadA-pdhR.* The product of the gene *dadA* found in alanine degradation pathway catalyzes the metabolic reaction forming the product 'pyruvate' which acts as a cofactor for the transcription factor PdhR.

4.4.2 Bow-tie Architecture in IMN

Connectivity analysis showed that the IMN was found to exhibit a bow-tie connectivity structure as previously shown for metabolic network of *E.coli* (Csete and Doyle 2004; Ma and Zeng 2003b). The bow-tie structure of IMN revealed a highly connected GSC (474 genes) having links from nodes in the IN subnetwork (94 genes) and links to the nodes in the OUT subnetwork (695 genes), TENDRILS (224 genes) originating from IN and OUT subnetworks, and TUBES (4 genes) connecting the IN and OUT without having any linkage to the GSC (Fig.2). Approximately 4% of the IMN comprising largely TRI and few MRs is completely disconnected from the entire bow-tie which forms the subnetwork OTHERS whereas the other five distinct subnetworks make up approximately 96% of the IMN (bow-tie). In the whole network, GSC is the largest component which contains 3928 interactions (75 MPIs, 3349 MRs and 504 TRs) between 474 genes which amounts to 31% of all the genes in the IMN and the other 14 strong components identified are distributed in other sub-networks. The other 14 strong components are significantly smaller in size containing only three to eight genes. An analysis of the GSC revealed that it contains six (*crp, rpoS, ihf, hns, fis & dnaA*) of the eleven global regulators identified in the previous work (Antonio *et al.* 2003; Ma *et al.* 2004b; Ma *et al*. 2004c; Rudd 2000) whereas the other five regulators (*cspA, rpoN, phoB, soxR & rpoE*) (Antonio *et al.* 2003; Mazurie *et al*. 2005; Babu and Teichmann 2003) are found in the IN subnetwork since these genes are involved only in the regulation of other genes in IN and GSC and there exists no feedback interaction toward them. Intriguingly all the FBL are found in the GSC but not in any one of the other 14 strong components since they are comprised of only metabolic interactions.

Further functional analysis revealed that the GSC contains many of the most conserved metabolic pathways of *E. coli* whose enzymes are universally present in the three domains (archaea, bacteria and eucarya) of life (Alvarez *et al*. 2003). Of the 23 most conserved metabolic pathways found by Peregrin-Alvarez *et al*, we identified nine pathways present completely whereas six others present partially with more than 60% of the genes occurring in the GSC (Table 2).

Table 2. The most conserved metabolic pathways in E. coli found in the giant strong component of the integrated molecular network.

Metabolic pathway	Superclass	Fraction of metabolic genes in GSC
Tryptophan biosynthesis	Individual amino acids	1.00
Nonoxidative branch of the pentose phosphate pathway	Energy-metabolism	1.00
Riboflavin, FMN and FAD biosynthesis	Cofactor-biosynthesis	1.00
Threonine biosynthesis	Individual amino acids	1.00
Ribose catabolism	Carbon-degradation	1.00
Methylglyoxal metabolism	Central-metabolism	1.00
Glyoxylate cycle	Energy-metabolism	1.00
TCA cycle, aerobic respiration	Energy-metabolism	1.00
FormylTHF biosynthesis	Intermediary-metabolism	0.84
Glycolysis	Energy-metabolism	0.71
Gluconeogenesis	Central-metabolism	0.70
Histidine biosynthesis	Individual amino acids	0.62
Isoleucine biosynthesis	Individual amino acids	0.60
Leucine biosynthesis	Individual amino acids	0.50

These pathways are involved in central intermediary metabolism, cofactor biosynthesis, energy metabolism, individual amino acid biosynthesis and sugar degradation. Many of them involve reactions catalyzed by the 11 most conserved metabolic enzymes according to their sequence identity to homologs in *Homo sapiens*. We searched for these enzymes in GSC. Intriguingly, our analysis found that GSC contains almost all the genes coding for the most conserved metabolic enzymes in *E. coli* (Alvarez *et al*. 2003). Among these eleven most conserved metabolic genes, ten of them (*guaC, gapA, sucD, pgi, fumC, mdh, galT, kbl, sdhA & gabD*) are identified in GSC and most of them are involved in energy metabolism. These facts support the idea that a bow-tie representation of IMN substantially reduces the complexity of analysis of this large scale network thereby yielding valuable insight into the subnetworks.

Six of the motifs (motifs 1, 2, 3, 5, 10 and 11) which make up the majority of the network motifs consist solely of metabolic reactions. The other seven motifs which include FFL and feedback regulations are still highly significant even though they represent only 17% of the total motifs. All five composite motifs are found mostly in GSC occurring together with the motifs composed entirely of MRs. Only one fifth of the FFLs are distributed in the GSC (Table. 2). Of the 391 target genes in 818 FFLs, 259 of them are found in the OUT subset and regulated by genes in the other subnetworks of the bow tie structure. This reveals that the majority of the FFLs are sparsely distributed across the network formed by the genes located in various subnetworks.

Chapter 5

Modularity Analysis of IMN

5.1 Introduction

Cellular functions are likely to be carried out in a highly modular manner. Modularity is shown to be the fundamental design attribute in the organization of robust and sustainable complex systems such as social networks, world wide web and engineered systems (barabasi and Oltvai 2004). A significant number of complex interactions between genes, proteins and metabolites which regulate the cellular processes are available from more reliable literature data (Ma *et al*. 2003a, 2004c; Salgado *et* al. 2004; Karp *et* al. 2002a; Wall *et* al. 2004). Most of the studies based on these data have predominantly studied the modular organization of molecular network comprising one particular type of interaction such as metabolic reaction, protein-protein interaction or gene expression (Ihmels *et* al. 2002; Segal *et al*. 2003; Ravasz *et al*. 2002). It has been recently reported that transcriptional regulatory, protein-protein interaction and metabolic networks of *E.coli* are scale-free network comprising topologic modules connected in a hierarchical way (Ravasz *et al*. 2002; Resendis-Antonio *et* al. 2005). To characterize completely the

various cellular processes and their regulation in order to have a more comprehensive understanding of such systems, it is relevant to have an integrative approach of analyzing diverse datasets. The integration of multiple molecular interactions into a single multi-scale network for functional studies has been addressed in several recent studies (Broder *et al*. 2000; Csete and Doyle 2004; Yeger-Lotem *et al*. 2004; Zhang *et al*. 2005). In particular, network motifs in an integrated network of *E. coli* (Yeger-Lotem *et al*. 2004) which was composed of transcriptional regulation interactions (TRIs) and protein-protein interactions (PPIs) were studied. However, the most important biological processes such as metabolic reactions and metabolite-protein interactions were not considered in this pioneer work (Yeger-Lotem *et al*. 2004). Here we have integrated a complex network from various molecular networks of *E. coli* comprising three types of interactions: TRIs between proteins and genes, metabolic reactions (MRs) catalyzed by proteins (enzymes) and metabolite-protein interactions (MPIs). Identifying the modular organization of such an integrated network using network decomposition methods can help us in better understanding the organization principle of complex systems. Various clustering methods can be used for network decomposition, for example simulated annealing, etc. To analyze this complex system, we have used modularity as a criterion for clustering since it is much simpler and faster for calculation. Here, we developed an algorithm to find a set of modules that comprise many within-module links and as few as possible between-module links. Application of our algorithm to such a multiple-interaction network spanning three different molecular levels efficiently grouped together genes of known similar function in well-defined functional modules. Additionally, we analyzed this multiple interaction network systematically to explore the network motifs involving three types of interactions. To discover these small network motifs, we applied the algorithm for identifying composite network motifs to the IMN data (Yeger-Lotem *et al*. 2004) which revealed 12 motifs and four among them are composite network motifs comprising multiple interaction types (Yeger-Lotem *et al*. 2004, Zhang *et al*. 2005). We examined the distribution of various motifs within and between the set of modules. Our analysis revealed that they connect several functional modules which are closely related to each other thereby supporting the fact that these smaller network motifs can be interpreted as basic building blocks of cellular circuitry (Yeger-Lotem *et al*. 2004).

Here, we have used modularity as a parameter of clustering and performed modular analysis on an integrated molecular network of *Escherichia coli* constructed from diverse collection of datasets involving metabolic reactions, metabolite protein interactions and transcriptional regulation. We have found that clustering this complex network significantly grouped together genes of known similar function in well-defined physiologically related modules. Identification of network motifs and correlating them with the modules of highly connected nodes may define their potential functional role. To this end, we detected and analyzed twelve highly significant three-node network motifs among which four are composite network motifs comprising multiple types of interactions. Distribution analysis of these motifs within and between the various functional modules supported the fact that these motifs represent basic patterns of regulation and organization of genes into modules. This study presents a basic framework for detecting functional modules and their interaction with various motifs in an integrated *E.coli* system.

5.2 Detecting Modules

Based on the modularity method of Newman and Girvan (Girvan and Newman 2002), a clustering tool is developed. In this algorithm, each node in the network is considered as a module itself, so at the beginning, the number of modules is the same as the number of nodes. Then, the modules are joined in pairs forming new modules, choosing at each step the join that results in the greatest increase (or smallest decrease) in modularity. The progress of the algorithm can be represented as a "Dendrogram", a module tree that shows the order of the joins and connects related modules (Fig. 15). Cuts through this dendrogram at different levels give divisions of the network into larger or smaller number of modules and the best cut can be chosen by using the maximal value of modularity as a criterion. To use modularity as a criterion for clustering, we calculated the change in modularity caused by the union of modules i and j (Girvan and Newman 2002):

$$\Delta M_{i,j} = \frac{l_{i,j}}{L} - \left(\frac{d_{i,j}}{2L} \right)^2$$

where $l_{i,j}$ is the number of links between the two modules, L is the total number of links in the network and $d_{i,j}$ is the sum of the degree of the nodes in the two modules.

5.2.1 Robustness of the Modularity Method

For each step in the calculation, the following multi-criteria approach is used:

1. First, select to cluster the modules which will increase more (or decrease less) the final modularity of the network, based on equation;
2. If more then one pair of nodes exist that fits on the first criterion, then select the ones with lower closeness centrality (modules at the periphery);
3. If there is still more then one pair of candidates for clustering, choose the one that will generate a new module with the smallest total degree;
4. If the last criterion selects more then one pair, choose the one which generates a new module with the smallest number of nodes;
5. Finally, if none of the above criteria select a single pair of nodes, select the first one from the generated list.

5.3 Using Modularity for Network Decomposition

The modularity analysis is performed on the integrated dataset of *E.coli* constructed from reliable literature based data containing metabolic reactions, metabolite protein interactions and transcriptional regulations. This network comprises 694 genes and all of these genes are functionally annotated (Kanehisa and Goto 2000). But we observed that using only modularity as criterion to choose the genes for clustering may not be very robust. This is because, particularly at the beginning of the clustering, many pairs of nodes contribute the same for the modularity and the clustering results vary depending on which pair is chosen. If, more than one pair of candidates for clustering exists, this problem can be minimized by choosing always the first pair and the result generated using this approach is mostly the same. But the order in which the genes appear in the list is probably based on the order they are stored in memory, which in turn is based on the order the nodes are read from an input data. In this case, changing the order of the nodes in the input data can lead to different results, decreasing the robustness of the algorithm. In order to increase the robustness of the modularity method, we used here a multi criteria approach. Applying this method to the integrated dataset clustered the entire network into smaller groups, generated a clustering tree that assembled all nodes into a single tree and finally dissected the tree into several modules by computing the best modularity.

Analyzing the modules of clustered genes revealed that genes sharing similar physiological function tend to group together.

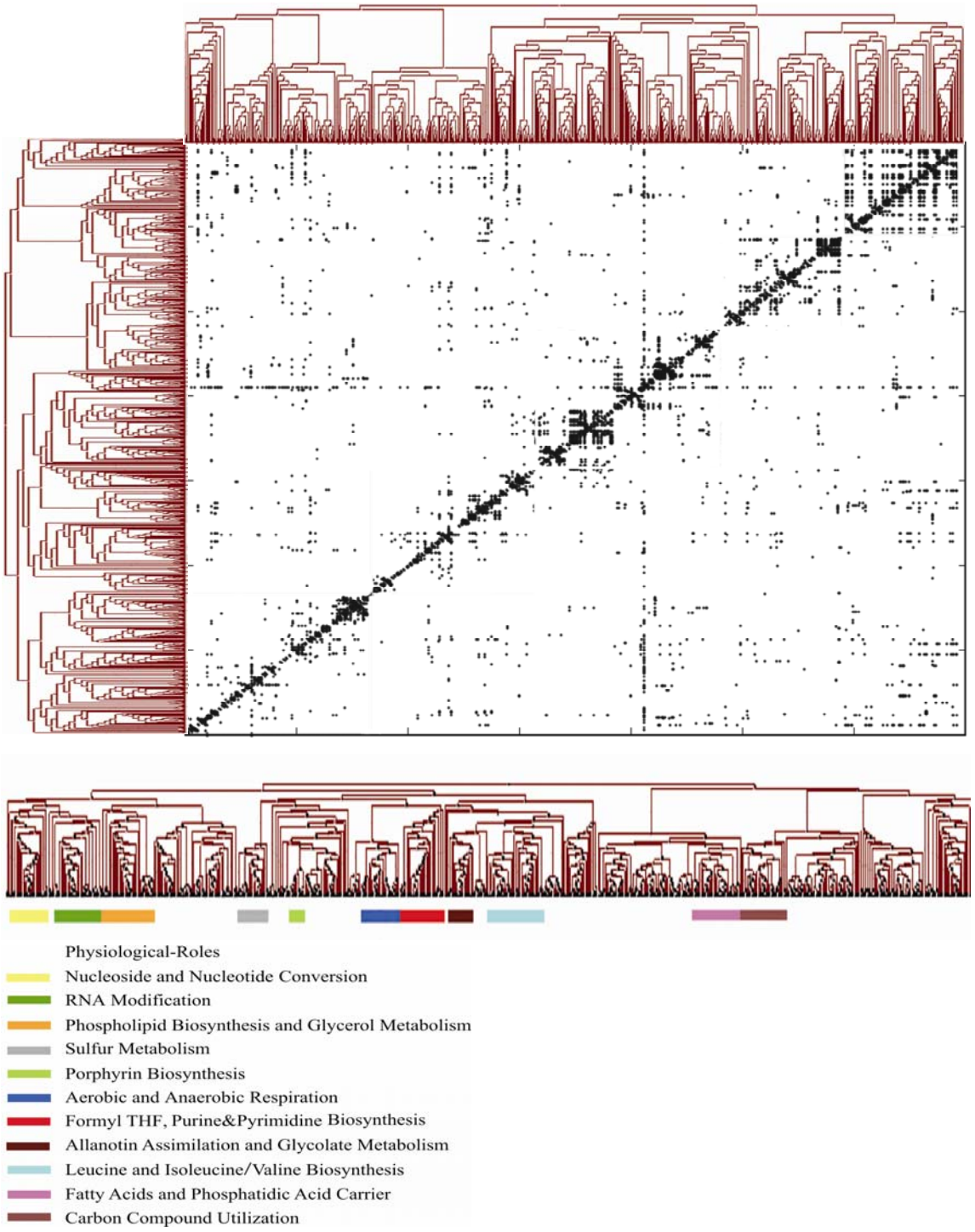


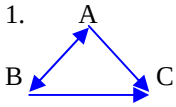
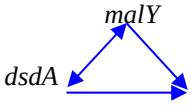
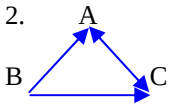
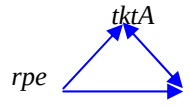
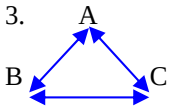
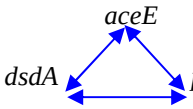
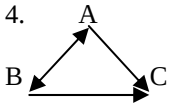
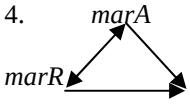
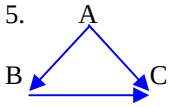
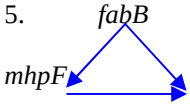
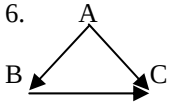
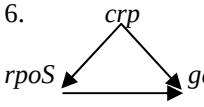
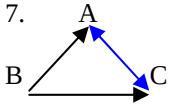
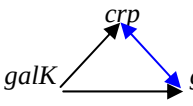
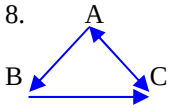
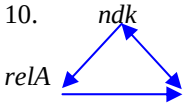
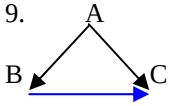
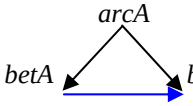
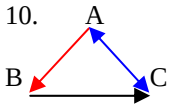
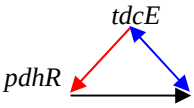
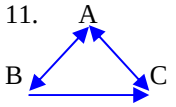
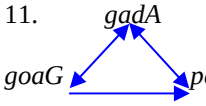
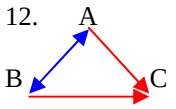
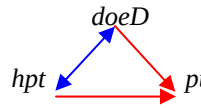
Figure 15. Clustered display of integrated molecular network of E.coli.
A. Each dot represents two genes connected to each other in the network and are placed in the same module. B. Dendrogram visualizing the various functional modules whose physiological roles are represented in various color codes.

For functional correlation we used the classification system for defining cellular or physiological roles of *E.coli* gene products, defined by Riley (Riley 1998). The various modules of genes which control specific biological responses are described in Figure 15. For example, the aerobic and anaerobic respiration module contains 11 out of 13 genes carrying out the similar physiological role. Majority of genes in this module synthesize enzymes which allow *E.coli* to use alternative compounds as electron acceptors thereby facilitating anaerobic growth. The global regulators 'aerobic respiration regulatory gene' (*arcA*) and 'fumurate and nitrate reductase regulatory gene' (*fnr*) are well clustered within the respiration module. The various functional modules which are clustered include: Fatty acid biosynthesis (M1); Allanotin assimilation and glycolate metabolism (M2); RNA modification (M3); Nucleoside and nucleotide conversion (M4); Phospholipid biosynthesis and glycerol metabolism (M5); Leucine and isoleucine/valine biosynthesis (M6); Formyl THF, purine and pyrimidine biosynthesis (M7); Sulfur metabolism (M8); Aerobic and anaerobic respiration (M9); porphyrin biosynthesis (M10) and carbon utilization (M11). Most of the modules are interconnected to each other and the linkage between them is influenced by specific significant patterns known as 'network motifs' which are overrepresented when compared to a randomized version of the real network. But there are some functional modules which are loosely connected to others. For example, there are only two interactions which connect the sulfur metabolism module to carbon metabolism and leucine and isoleucine/valine biosynthesis modules. This indicates that analyzing the distribution of the network motifs may shed more light in understanding how these modules and the interactions between them are interplayed.

5.4 Motif Distribution among Modules

Network motifs are defined as patterns of interconnections that recur in many different parts of a network at frequencies much higher than those found in randomized networks (Shen-Orr *et al.* 2002). In analyzing the modules of IMN for three-gene substructures, we have also searched for composite network motifs (Yeger-Lotem *et al.* 2004): significant three-gene connected patterns comprising multiple interactions like transcriptional regulation, metabolic reaction and metabolite-protein interaction. Our analysis revealed 12 highly significant motifs (Table 3).

Table 3. Three-gene network motifs in the IMN of E. coli

Motif pattern*	Motif Example	Zscore†	Nreal	RndAvg	RndStd
1. 		204.1	2220	92	10.4
2. 		179.7	625	10	3.4
3. 		176.4	3981	365	20.5
4. 	4. 	25.7	50	2.1	1.9
5. 	5. 	24.8	344	90	10.2
6. 	6. 	12.4	285	86	16.1
7. 		11.5	280	107	15
8. 	10. 	11.4	100	31	6
9. 		10.8	219	92	11.8
10. 		10.4	21	2.7	1.8
11. 	11. 	9.9	272	132	14.2
12. 		5.1	13	3.2	1.9

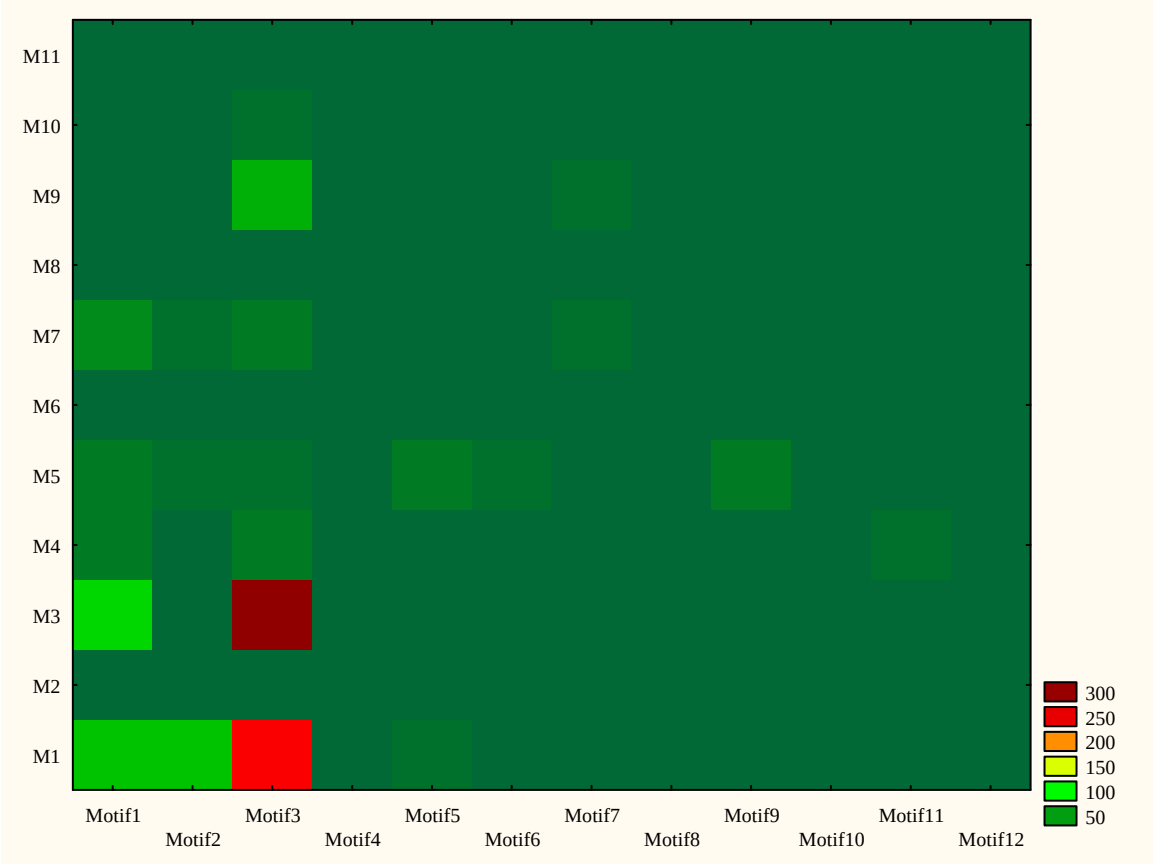
*Statistically significant motifs involving multiple interaction types represented by the edges: red represents Metabolite-Protein Interaction, blue represents Metabolic Reaction and black represents Transcription Regulaltion respectively.

†Z-score= (Nreal - RndAvg)/ RndStd where Nreal is the number of motifs occurring in the real network.

This included four composite motifs whereas the rest of the structures are composed of either MR (6 motifs) or TRI (2 motifs). The various motifs identified in the IMN are described below:

Motif 1, 2, 3, 5, 8 & 11, composed entirely of MRs are the most abundant single interaction type motifs in the IMN. Motif 4 includes a gene pair regulating each other that co-regulates a target gene. This motif suggests that the gene pair forming the loop has the tendency to belong to the same operon or code for transcription factors which function as a protein complex. Five pairs of coregulating genes are found and among these *glnG-glnL, gutM-srlR, marA-marR* and *rhaR-rhaS* are located in the same operon. Motif 6 is the transcriptional feed-forward (FF) motif comprising only TRIs in which one of the gene regulates the other and these two genes together regulate a target gene. Motifs 7 and 9 represent a pair of co-regulated genes whose protein products (enzymes) catalyze successive metabolic reactions whereas in the case of motif 9, the successive MRs are irreversible. Motif 10 comprises two genes catalyzing successive metabolic reactions whose final product acts as a cofactor of the third gene which in turn controls the transcriptional regulation of one of the metabolite genes forming a 'feedback loop'. Motif 12 comprises reversible successive metabolic reactions producing the same metabolite as the end product that feedbacks through metabolite-protein interaction. Although this motif is detected as significant in the IMN, it became insignificant in the functional modules. This motif is not found distributed within or between any of the functional modules but it connects the genes which are not clustered in any of the modules to the carbon metabolism module through MPIs.

It has been previously reported that network motifs demonstrate in detail the precise structure of the biological complexity of the regulatory network (Resendis-Antonio *et al.* 2005). Here we have analyzed for the distribution of all the 12 significant motifs within the various identified modules (Fig 16.). Almost 38% of the motifs are distributed within the same modules. From the 62% of the motifs which are distributed between various modules, 42% connect two different modules whereas the rest forms a connection between three different modules.



Module	Physiological-Roles
M1	Fatty acid biosynthesis
M2	Allanotin assimilation and Glycolate metabolism
M3	RNA Modification
M4	Nucleoside and nucleotide conversion
M5	Phospholipid Biosynthesis and Glycerol
M6	Leucine/isoleucine and Valine Biosynthesis
M7	Formyl THF, Purine and Pyrimidine Biosynthesis
M8	Sulfur Metabolism
M9	Aerobic and Anaerobic Respiration
M10	Porphyrin Biosynthesis
M11	Carbon Compound Utilization

Figure 16. A matrix representation of the distribution of detected motifs within the 12 identified modules.

Distribution analysis showed that 40% of most abundant single interaction type motifs composed entirely of MRs: Motif 1, 2, 3, 5, 8 & 11 are distributed within the same modules. These motifs often overlap with each other forming a homologous motif cluster which is a general property of most real networks (Barabasi and Oltvai 2004). It is noticed that they also overlap with composite motifs forming motif themes (Zhang *et al.* 2005) which are recurring higher-order interconnection pattern that constitutes multiple occurrences of network motifs. These structures represent the underlying biological phenomena and fundamental network design principles. For example, the module which clusters together the genes involved in leucine and isoleucine/valine biosynthesis comprises the motifs 1, 2, 3, 7, 9 and 11 (Figure 17). Motifs 7 and 9 are composite network motifs comprising TRIs and MRs whereas motifs 1, 2, 3, 11 comprise only MRs. All these motifs overlap with one another forming a motif theme which includes the genes displaying similar physiological functions. Motif 4 forms a connection between the genes that are well clustered within the carbon metabolism module. For example, the gene pair *gutM-srlR* which regulates each other forming a loop coregulates three other genes (*srlB, srlD and srlE*) having the same physiological function as *gutM-srlR* and they are all well placed in the carbon metabolism module. Analyzing the distribution of FF motif which has been previously well studied (Strohman 1997; Shen-Orr *et al.* 2002; Mangan and Alon 2003a; Ma *et al.* 2004c) revealed that only 24% of this motif is distributed within the modules such as carbon metabolism and respiration modules and the rest form a connection between other modules since most of the regulated target genes are found in various modules. Almost 40% of this motif connecting various modules involves the global regulator cyclic AMP receptor (*crp*) which is well located in the carbon metabolism module. Similarly 30% of the connections formed by network motifs 7 and 9 comprise the global regulator *crp* as the co-regulator gene which leads to the highest degree of overlap of these motifs with FF motif. For example, the gene *crp* regulates the two genes *sucA* and *sucB* which are involved in the TCA cycle forming the motif 7. Simultaneously *crp* regulates the global regulator *rpoS* and these two genes together regulate the target gene *sucA* thereby forming a connection between the motif 7 and FF motif. Motif 10, a feedback loop is carried out at the post-transcriptional level which plays an important role in the *E.coli* gene regulation.

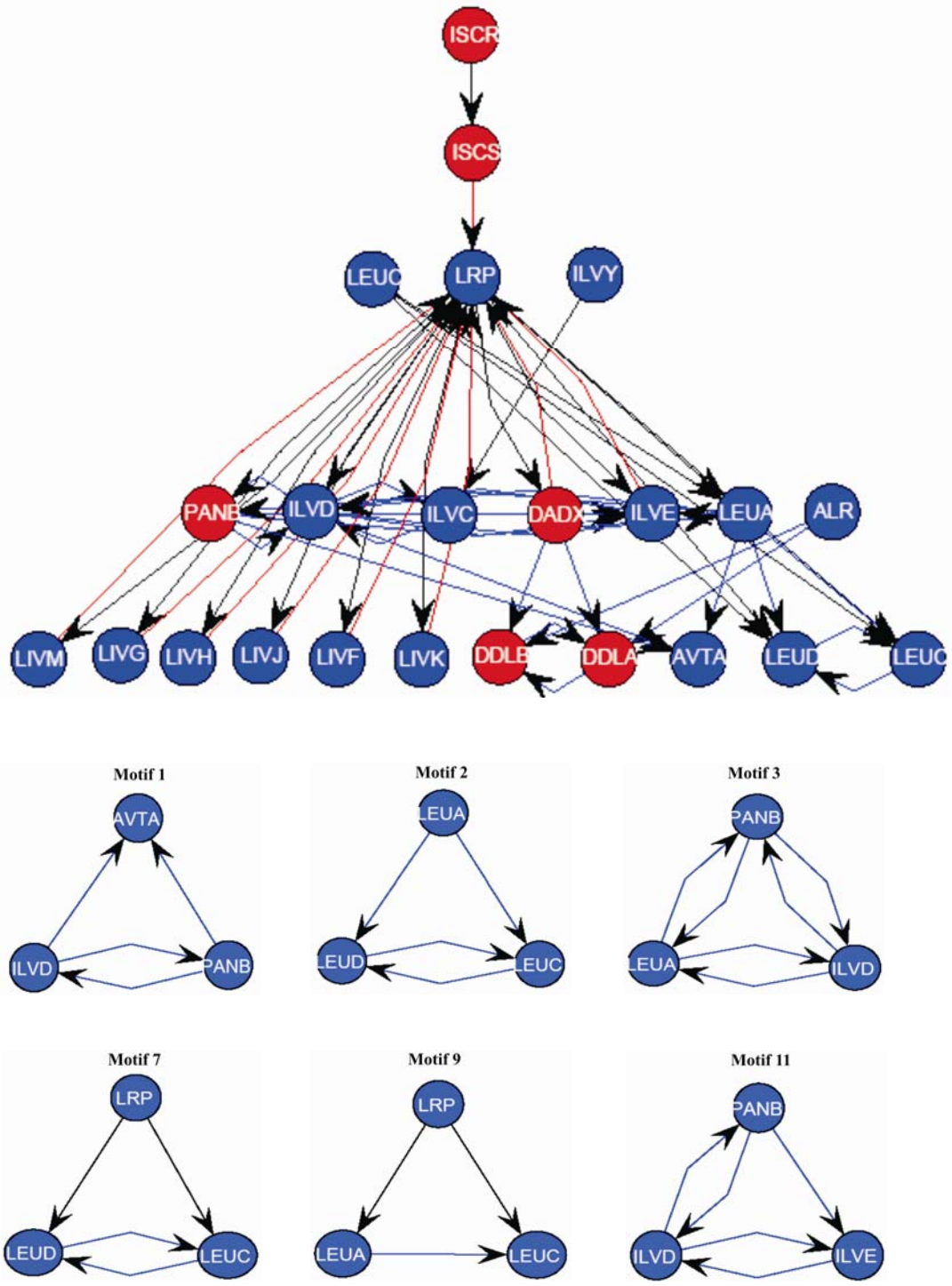
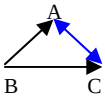
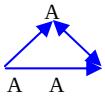
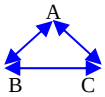
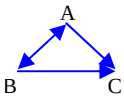
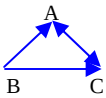
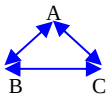
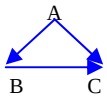
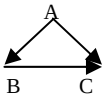
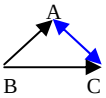
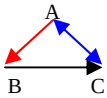
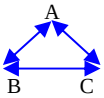
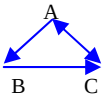
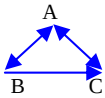
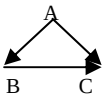
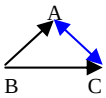
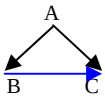
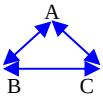
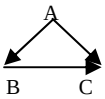
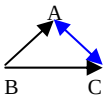
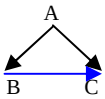
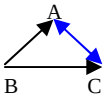
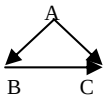
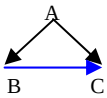
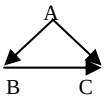


Figure 17. Distribution of motifs within the modules.

A. Hierarchical organization of Leucine and isoleucine/valine biosynthesis module. Blue nodes are genes involved in Leucine and isoleucine/valine biosynthesis, blue edge represents MR, black edge represent TRI and red edge represent MPI. B. Six different motifs which are found in the Leucine and isoleucine/valine biosynthesis module.

Twenty one feedback loops are identified from the motif analysis and 52% of this motif is found within the carbon metabolism module. We have also studied the relationship between motifs and the various modules connected by them (Table 4). Modules M1 and M7 are connected to each other by motifs 2 and 3 which makes sense since the fatty acid biosynthesis and formyl THF biosynthesis are linked through threonine metabolism. Motif 7 shows the relation between the fatty acid biosynthesis (M1) and carbon metabolism (M11) modules. Modules M6 and M7 are connected by the composite motif 9. FF motifs connect carbon metabolism and respiration modules which are closely related to each other through tricarboxylic acids cycles (TCA) (Resendis-Antonio *et al.* 2005). Moreover it also joins the modules M1 and M5 which are tightly coupled to each other since controlling the production of acyl moieties an early step in fatty acid biosynthesis most probably regulates phospholipid biosynthesis (Nunn *et al.* 1977; Rock and Jackowski 1982). A list of motifs connecting various modules is described in the supplementary material online. These instances explain that several single interaction type and composite motifs are involved in building the modules and also in establishing the connection between them.

Table 4. Motifs connecting the functional modules.

Functional Modules	Motifs
1. M1 and M5	 Motif 7
2. M1 and M7	 Motif 2  Motif 3
3. M1 and M11	 Motif 1  Motif 2  Motif 3  Motif 5  Motif 6  Motif 7  Motif 10
4. M3 and M4	 Motif 3  Motif 8  Motif 11
5. M3 and M5	 Motif 6  Motif 7  Motif 9
6. M3 and M7	 Motif 3
7. M5 and M11	 Motif 6  Motif 7  Motif 9
8. M6 and M7	 Motif 7
9. M1, M5 and M11	 Motif 6
10. M2, M6 and M7	 Motif 9
11. M3, M4 and M5	 Motif 6

Chapter 6

Discussions and Conclusions

6.1 Extended TRN

A more complete TRN of E.coli is generated by combining information from three different data sources (RegulonDB, Ecocyc and TRN-SO) and literature survey. Only a relatively small part of the regulatory interactions is found to be common in all the three datasets, indicating the importance of data integration for obtaining a more complete and reliable network. Structural analysis of the extended network reveals both the consistency and inconsistency found within results obtained from the network of Shen-Orr *et al.* (Shen-Orr *et al.* 2002) recently used in several studies. The new network preserves the multi-layer hierarchical structure but has more layers because of more interactions inside the network. FFLs still remain the only three-node network motif in the network but the number of FFLs increases greatly with the size of the network. The distribution of the different types of FFLs is also different from that derived from the TRN-SO network. Most of the FFLs are connected to form a giant motif cluster instead of forming several small disconnected clusters. Furthermore, only a small portion of the genes is solely regulated by only one FFL. Many of the genes are regulated by two or more interacting FFLs or other more complicated network motifs together with transcriptional factors not

belonging to any network motifs. These results underline the importance of having a more complete and reliable network in the study of structure and function of transcriptional regulation of gene expression.

6.2 Role of Feedback Regulations in Cellular Mechanism

Feedback interactions play a key role in executing and regulating cellular functions such as cell cycle and dynamic response to fluctuations of environmental conditions. Identifying such mechanisms is thus highly relevant for systems biology study of cells. In our previous work of the transcriptional regulatory network of *E. coli* (Ma *et al.* 2004c), we were surprised to find that there is almost no FBL at the transcriptional level. This prompted us to systematically examine the possible involvement of metabolic signals in feedback circuits by using an integrated molecular network. Such feedback circuits comprising metabolite-protein interaction could allow the cell to monitor the presence and level of such metabolites continuously and to fast respond at transcriptional level. We identified 69 such circuits by integrating MPI with TRN and MR networks. Feedback regulation circuits mediated through metabolite-protein interaction are also widely observed in other organisms. For example, it controls the oxidative-stress response in eukaryotes (Maeta *et al.* 2004). In *Saccharomyces cerevisiae*, *gre2* the transcription of which is positively regulated by *yap1* codes for an enzyme D-lactaldehyde dehydrogenase which catalyzes the conversion of lactaldehyde to methylglyoxal; the latter modulates the activity of the Yap1 transcription factor which is critical for the oxidative-stress response (Maeta *et al.* 2004). Preliminary study of the literature based transcriptional regulatory network of *S. cerevisiae* showed that it also has a multi-layer hierarchical structure similar with that of *E. coli*. Actually only one feed back loop (between *GZF3* and *DAL80*) at transcriptional level is found. This may imply that the feedback regulation of gene expression is generally mainly mediated by metabolite-protein interactions as found here for *E. coli*. It would be of great interest to examine to which extent it is prevailing in other organisms and what are the evolutionary driving forces for this mechanism.

6.3 Self-similar Structure of IMN

To simplify the structural analysis we first investigated the global structure of the integrated network. We investigated the self-similar structure (Dill *et al.* 2001; Pennock *et al.* 2002; Donato *et al.* 2005) of the IMN which exhibits the bow-tie structure repeating itself in the IN and OUT subnetworks. A self-similar object is exactly or approximately similar to a part of itself, i.e., the whole has the same shape as one or more of the parts.

A compact topological space *X* is self-similar if there exists a finite set *S* indexing a set of non-surjective homeomorphisms $\{f_s\}_{s\in S}$ for which

$$X = \cup_{s\in S} f_s(X)$$

If $X \subset Y$, we call *X* self-similar if it is the only non-empty subset of *Y* such that the equation above holds for $\{f_s\}_{s\in S}$. We call

$$\mathfrak{L} = (X, S, \{f_s\}_{s\in S})$$

a self-similar structure.

Many objects in the real world, such as coastlines, are statistically self-similar: parts of them show the same statistical properties at many scales. Self-similarity is a typical property of fractals. Self-similarity also has important consequences for the design of computer networks, as typical network traffic has self-similar properties.

Provided that the in-degree, out-degree and Strong core component (SCC) size distributions in the IN and OUT components are the same for the complete IMN, it is tempting to conjecture that the IMN has a self-similar structure, i.e., the bow-tie repeats itself inside the IN and OUT components. For this purpose we examined the distribution of the strong components and the weakly connected components (WCCs). The largest strong components that are distributed in the IN and OUT subnetworks are composed of only three and eight nodes respectively which are minuscule compared with the GSC in the IMN. This fact reveals that there is no large strong component in the IN and OUT

subnetworks which can form a GSC in a nested bow-tie structure. This is a first indication that the IMN doesn't possess a self-similarity property unlike web graph since there exists no sizable SCC in the IN and OUT components that could play the role of the giant strong component (GSC) in a potential bow-tie structure. But it is still possible that there exist s a giant weakly connected component (WCC) within IN and OUT. Therefore, we searched for WCCs in the subnetworks and computed 45 (48% of interactions) and 534 (77% of interactions) of them in the IN and OUT subnetworks respectively. These WCCs are large in number but the largest WCC in the IN subnetwork consists of only 14% of its nodes, whereas the largest WCC in the OUT subnetwork consists of 11% of its nodes. The largest WCCs of both of these components are insignificant, compared with the one detected in the IMN which contains 96% of the nodes. These findings support the hypothesis that the self-similarity conjecture may not be true for cellular networks (Donato *et al.* 2005).

6.4 Function of Motifs in IMN

Network motifs are considered as a fundamental characteristic of biological systems (Lee *et al.* 2002; Shen-Orr *et al.* 2002). Previous studies have been mainly focused on motifs of networks composed of one type of molecules. Using the integrated network of *E. coli* we studied the prevalence of motifs spanning different molecular levels. This study identified 13 significantly enriched network motifs in the IMN of *E. coli*. Six of them (Table 2; Motif 1, 2, 3, 5, 10 & 11) are composed only of metabolic reactions. Abundance of such subgraphs increases the extent of clustering and modularity of the network. Motif 4 represents two genes regulating each other, thereby forming a loop that co-regulates a target gene. This motif suggests that gene pairs regulating each other and further co-regulating a target gene have the tendency to belong to the same operon or code for transcription factors which function as a protein complex. Motif 6 is the transcriptional feed-forward loop. The different types of this motif (coherent and incoherent FFLs) were studied previously and showed that the coherent FFL speeds up the responses of the target gene while the incoherent FFL delays the response. Motifs 7 and 8 represent a pair of coregulated genes whose protein products (enzymes) catalyze successive metabolic reactions. These motifs are found in several cellular pathways and

in most cases the coregulated genes occurring in a specific pathway are organized in the same operon. Taken both the above facts into consideration, it suggests that most of the coregulated metabolite genes should be involved in the same metabolic pathway. This motif also suggests that some metabolite genes catalyzing the successive metabolic reactions should be coregulated at the transcriptional level. Motif 9 is a feedback loop with two genes catalyzing successive metabolic reactions whose final product acts a cofactor of the third gene which in turn controls the transcriptional regulation of one of the metabolite genes. These feedback loops are carried out at the post-transcriptional level which plays an important role in the *E.coli* gene regulation. Motif 12 represents reversible successive metabolic reactions producing the same metabolite as the end product that feedbacks through metabolite-protein interaction. It's the lowest occurring significant motif in the network due to the limitation of the interaction data between metabolites and proteins. Motif 13 comprises TRI and MPI and of the 24 occurrences of this motif, there are 19 cases where there exist feedback regulations to *pdhR* through MPI. It's because pyruvate that acts as a cofactor for *pdhR* is produced from many enzymatic reactions catalyzed by the metabolic genes which are regulated by the global regulator *crp*. This motif is almost similar to feed-forward loop, but one of the TRIs is replaced by a MPI since the network comprises multiple interaction types which makes it different from FFL.

A more comprehensive study is needed to combine multivariate metabolomic and proteomic data to identify metabolite-protein correlation. The present approach can be further extended with other interaction types like PPI and sRNA interactions in order to study the cellular process more precisely. Moreover, clustering the IMN into biologically meaningful modules might provide deeper understanding of structural and functional properties of the IMN.

The interaction network of *E. coli* established in this work by integrating data on metabolic reactions, metabolic-protein interactions and transcriptional regulations represents a necessary and useful step towards a system level understanding of cellular processes and their regulation. By analyzing the integrated network we found that feedback regulation of gene expression in *E. coli* is primarily mediated by metabolite-protein interaction. 69 such feedback regulation loops of gene expression were identified

which are important to understand the system behavior of cells. We further showed the existence of a global bow-tie architecture which spans three molecular levels. The bow-tie structure was previously found in metabolic networks, TLR-mediated innate immune system, transcriptional and translational process and the webgraph (Bategelj and Mrvar 1998; Broder *et al.* 2000; Gutierrez-Rios *et al.* 2003; Ma and Zeng 2003b; Ma *et al.* 2004a; Donato *et al.* 2005). The ubiquity of bow-tie structure in all these networks indicates that this architecture may represent a fundamental organizing principle for robust technical and biological systems. Based on the integrated network and the global bow-tie structure we detected thirteen highly significant network motifs comprising five composite network motifs which include the FBL mediated through MPI. Most of the network motifs overlap together and thereby form the GSC of the bow-tie. Future studies could aim at analyzing the IMN from broader perspectives by adding other interaction types which relates two genes and examining the dynamics of interactions by incorporating high-through-put experimental data.

The study uncovered the presence of modularity in the IMN of *E.coli* and revealed several modules assigned with genes characterized by similar physiological functions. Using a multi-criteria approach, we improved the robustness of the modularity method. Further we increased the speed of the modularity algorithm by more than 20 times depending on the size of the network. Recent studies have provided insight into the composite network motifs of the molecular networks by presenting approach that can be used to analyze multiple interaction type networks (Yeger-Lotem *et al.* 2004). This approach is used here to examine the distribution of patterns of interconnection within and between the identified modules comprising physiologically related genes. Analysis revealed that the network motifs distributed between different modules are bit higher than those found within the modules. Overlapping of several motifs resulted in the formation of network themes that may reduce the complexity of interpreting most of the network motifs and also in simplification of the complex integrated network. It would be interesting to analyze the modularity of a more detailed network of *E.coli* by adding more biological interaction types which may reveal more complex relationships. Moreover the same application could be extended to the integrated network of other organisms as well

Appendix A

Abbreviations

Clu	**Cluster**
$d_{i,j}$	**Sum of the degree of two nodes in two modules**
EG	**Ecogene**
FBL	**Feedback loop**
FF	**Feed-forward**
FFL	**Feed-forward loop**
GML	**Graph modeling language**
GSC	**Giant strong component**
HTML	**Hyper text markup language**
IMN	**Integrated molecular network**
JRE	**Java runtime environment**

$l_{i,j}$	**Total number of links in the network**
L	**Total number of links in the network**
$\Delta M_{i,j}$	**Change in modularity**
MPI	**Metabolite-protein interaction**
MR	**Metabolic reaction**
na	**Node attributes**
NET	**Network**
Nreal	**Number of motifs in the real network**
P	**Probability**
PPI	**Protein-protein interaction**
RNA	**Ribonucleic acid**
RndAvg	**Random average**
RndStd	**Random standard deviation**
SIF	**Simple interactions format**
SC	**Strong component**
TCA	**Tricarboxylic acid**
TRI	**Transcription regulation interaction**
TRN	**Transcriptional regulatory network**
txt	**Text**

WCC **Weakly connected component**

Appendix B

Links

Systems Biology group at GBF

http://www.gbf.de/SystemsBiology

Cytoscape

http://www.Cytoscape.org

Ecocyc

http://www.ecocyc.org

EcoTFs

http://ecotfs.lanl.gov/table1.html

GenProtEC

http://genprotec.mbl.edu/

GML file format

http://www.infosun.fmi.uni-passau.de/Graphlet/GML/

Human genome project

http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

Java language

http://java.sun.com

Mfinder 1.2

http://www.weizmann.ac.il/mcb/UriAlon/

Network conversion tool

http://pynetconv.sourceforge.net

NetworkX

http://networkx.sourceforge.net/

Pajek

http://vlado.fmf.uni-lj.si/pub/networks/pajek/

Python language

http://python.org/

RegulonDB

http://cifn.unam.mx/Computational_Genomics/regulondb/

Appendix C

Mfinder Results

MOTIF FINDER RESULTS:

Network name: IN_Motif_analysis_MPI.txt

Num of Vertices: 1549 Num of arcs: 6497 Num of colors: 3

Num of Connected Vertices: 0

Motif size searched is 3

Number of random network generated is: 1000

Algorithm used: Exhaustive Search

Distribution of arc profiles (edges with arcs both ways are counted in two profiles)

Arc Profile frequency

0 (0,0) 0

1 (0,1)	0
2 (0,2)	0
3 (0,3)	0
4 (0,4)	0
5 (0,5)	0
6 (0,6)	0
7 (0,7)	0
8 (1,0)	1078
9 (1,1)	2688
10 (1,2)	0
11 (1,3)	0
12 (1,4)	0
13 (1,5)	0
14 (1,6)	0
15 (1,7)	0
16 (2,0)	2591
17 (2,1)	0
18 (2,2)	14
19 (2,3)	0
20 (2,4)	31
21 (2,5)	0
22 (2,6)	0
23 (2,7)	0

24 (3,0)	0
25 (3,1)	0
26 (3,2)	0
27 (3,3)	0
28 (3,4)	0
29 (3,5)	0
30 (3,6)	0
31 (3,7)	0
32 (4,0)	64
33 (4,1)	0
34 (4,2)	31
35 (4,3)	0
36 (4,4)	0
37 (4,5)	0
38 (4,6)	0
39 (4,7)	0
40 (5,0)	0
41 (5,1)	0
42 (5,2)	0
43 (5,3)	0
44 (5,4)	0
45 (5,5)	0
46 (5,6)	0

47 (5,7) 0

48 (6,0) 0

49 (6,1) 0

50 (6,2) 0

51 (6,3) 0

52 (6,4) 0

53 (6,5) 0

54 (6,6) 0

55 (6,7) 0

56 (7,0) 0

57 (7,1) 0

58 (7,2) 0

59 (7,3) 0

60 (7,4) 0

61 (7,5) 0

62 (7,6) 0

63 (7,7) 0

Motif frequency table

Note that self isomorphic motifs are counted more than once

Real	RndAvg	RndStD	Zscore	Pvalue	Motif	Motif ID
1424	2294	10.4	-84.0	1.0000	| 0 1 1 | 0 0 0 | 0 0 0 |	72
3	3.0	0.2	0.2	0.9750	| 0 2 1 | 0 0 0 | 0 0 0 |	80
73	79	1.2	-4.9	1.0000	| 0 4 1 | 0 0 0 | 0 0 0 |	96

118366	119300	38.9	-24.0	1.0000	| 0 2 2	| 0 0 0	| 0 0 0 |	144
4	3.9	0.2	0.2	0.9530	| 0 4 4	| 0 0 0	| 0 0 0 |	288
2587	2902	14.4	-21.9	1.0000	| 0 0 1	| 1 0 0	| 0 0 0 |	576
3695	8229	25.4	-178.4	1.0000	| 0 1 1	| 1 0 0	| 0 0 0 |	584
10	9.9	0.3	0.3	0.9250	| 0 0 2	| 1 0 0	| 0 0 0 |	640
50	49	1.1	1.1	0.3230	| 0 0 4	| 1 0 0	| 0 0 0 |	768
216	256	3.8	-10.5	1.0000	| 0 1 4	| 1 0 0	| 0 0 0 |	776
1415	1568	10.6	-14.4	1.0000	| 0 0 1	| 2 0 0	| 0 0 0 |	1088
10	9.8	0.4	0.4	0.8370	| 0 4 1	| 2 0 0	| 0 0 0 |	1120
2229	2793	34.8	-16.2	1.0000	| 0 0 2	| 2 0 0	| 0 0 0 |	1152
37	426	13.1	-29.7	1.0000	| 0 2 2	| 2 0 0	| 0 0 0 |	1168
104	122	4.1	-4.4	1.0000	| 0 0 4	| 2 0 0	| 0 0 0 |	1280
2	1.9	0.2	0.2	0.9470	| 0 4 4	| 2 0 0	| 0 0 0 |	1312
471	496	1.7	-14.8	1.0000	| 0 0 2	| 4 0 0	| 0 0 0 |	2176
851	860	1.7	-5.5	1.0000	| 0 2 2	| 4 0 0	| 0 0 0 |	2192
1370	3787	14.3	-169.2	1.0000	| 0 0 1	| 0 0 1	| 0 0 0 |	32832
344	89	9.7	26.3	0.0000	| 0 1 1	| 0 0 1	| 0 0 0 |	32840
1199	1348	10.5	-14.2	1.0000	| 0 0 2	| 0 0 1	| 0 0 0 |	32896
221	74	10.5	14.0	0.0000	| 0 2 2	| 0 0 1	| 0 0 0 |	32912
2256	94	10.8	201.0	0.0000	| 0 1 1	| 1 0 1	| 0 0 0 |	33352
3	1.2	1.0	1.8	0.0020	| 0 0 2	| 4 0 1	| 0 0 0 |	34944
1964	2720	35.6	-21.2	1.0000	| 0 0 2	| 0 0 2	| 0 0 0 |	65664
818	257	34.9	16.1	0.0000	| 0 2 2	| 0 0 2	| 0 0 0 |	65680

89	104	4.0	-3.8	1.0000	| 0 0 4	| 0 0 2	| 0 0 0 |	65792
199	5.2	6.6	29.5	0.0000	| 0 2 2	| 2 0 2	| 0 0 0 |	66704
24	8.6	4.0	3.8	0.0000	| 0 0 4	| 2 0 2	| 0 0 0 |	66816
506	515	2.2	-4.2	0.9990	| 0 0 4	| 0 0 4	| 0 0 0 |	131328
13	3.2	1.9	5.2	0.0000	| 0 1 4	| 1 0 4	| 0 0 0 |	131848
2086	3526	16.2	-88.8	1.0000	| 0 1 0	| 1 0 0	| 1 0 0 |	262664
5286	16296	61.1	-180.1	1.0000	| 0 1 1	| 1 0 0	| 1 0 0 |	262728
39	49	1.2	-8.6	1.0000	| 0 4 0	| 2 0 0	| 1 0 0 |	263200
3131	3541	24.8	-16.5	1.0000	| 0 0 1	| 2 0 0	| 1 0 0 |	263232
38	41	1.0	-3.1	0.9940	| 0 4 1	| 2 0 0	| 1 0 0 |	263264
4	7.1	2.8	-1.1	0.9050	| 0 1 0	| 0 0 1	| 1 0 0 |	294920
100	31	6.1	11.2	0.0000	| 0 1 1	| 0 0 1	| 1 0 0 |	294984
625	10	3.3	186.1	0.0000	| 0 0 1	| 1 0 1	| 1 0 0 |	295488
272	131	13.9	10.1	0.0000	| 0 1 1	| 1 0 1	| 1 0 0 |	295496
21	1.8	1.4	13.8	0.0000	| 0 4 1	| 0 0 2	| 1 0 0 |	327776
7	1.7	1.4	3.8	0.0000	| 0 4 0	| 2 0 2	| 1 0 0 |	328736
280	86	12.3	15.7	0.0000	| 0 0 1	| 2 0 2	| 1 0 0 |	328768
2	1.5	1.3	0.4	0.1560	| 0 4 1	| 2 0 2	| 1 0 0 |	328800
10	19	0.8	-11.7	1.0000	| 0 4 0	| 2 0 0	| 2 0 0 |	525344
7	15	0.6	-12.3	1.0000	| 0 2 0	| 4 0 0	| 2 0 0 |	526352
3	1.5	1.3	1.2	0.0290	| 0 4 0	| 0 0 2	| 2 0 0 |	589856
2	1.1	1.0	0.8	0.0320	| 0 4 0	| 2 0 2	| 2 0 0 |	590880
9	1.2	1.0	8.1	0.0000	| 0 0 2	| 2 0 2	| 2 0 0 |	590976

8	1.2	1.0	6.7	0.0000	| 0 0 4	| 2 0 2	| 2 0 0 |	591104
55	59	0.6	-6.4	1.0000	| 0 0 4	| 0 0 4	| 2 0 0 |	655616
4	1.1	1.1	2.6	0.0000	| 0 0 4	| 1 0 4	| 2 0 0 |	656128
39	40	0.2	-3.8	0.9970	| 0 2 2	| 4 0 0	| 4 0 0 |	1050768
3982	359	19.6	184.5	0.0000	| 0 1 1	| 1 0 1	| 1 1 0 |	2392648
1	1.0	1.0	0.0	0.0210	| 0 2 2	| 4 0 1	| 4 1 0 |	3180688

The following 'significant' motifs were found (out of 55 motifs altogether):

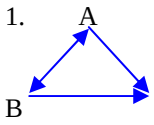
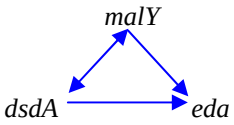
Real	RndAvg	RndStD	Zscore	Pvalue	Motif			Motif ID
344	89	9.7	26.3	0	| 0 1 1	| 0 0 1	| 0 0 0 |	32840
221	74	10.5	14	0	| 0 2 2	| 0 0 1	| 0 0 0 |	32912
2256	94	10.8	201	0	| 0 1 1	| 1 0 1	| 0 0 0 |	33352
818	257	34.9	16.1	0	| 0 2 2	| 0 0 2	| 0 0 0 |	65680
199	5.2	6.6	29.5	0	| 0 2 2	| 2 0 2	| 0 0 0 |	66704
24	8.6	4	3.8	0	| 0 0 4	| 2 0 2	| 0 0 0 |	66816
13	3.2	1.9	5.2	0	| 0 1 4	| 1 0 4	| 0 0 0 |	131848
100	31	6.1	11.2	0	| 0 1 1	| 0 0 1	| 1 0 0 |	294984
625	10	3.3	186.1	0	| 0 0 1	| 1 0 1	| 1 0 0 |	295488
272	131	13.9	10.1	0	| 0 1 1	| 1 0 1	| 1 0 0 |	295496
21	1.8	1.4	13.8	0	| 0 4 1	| 0 0 2	| 1 0 0 |	327776
280	86	12.3	15.7	0	| 0 0 1	| 2 0 2	| 1 0 0 |	328768
3982	359	19.6	184.5	0	| 0 1 1	| 1 0 1	| 1 1 0 |	2392648

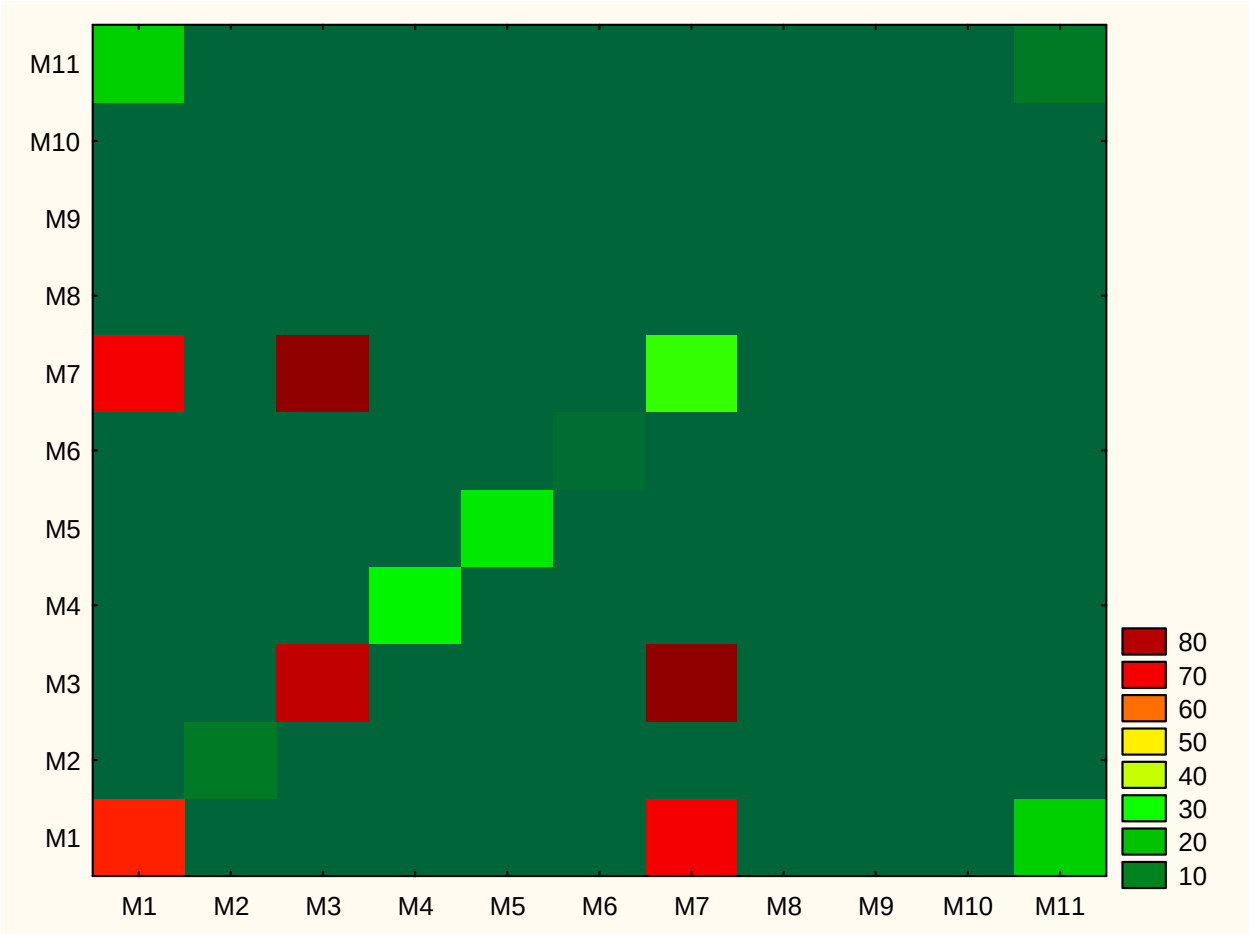
Appendix D

Motif Distribution

In this appendix, the matrix representations of all the identified motifs formed by the interaction of every specific pair of modules are shown. (Section 5.4)

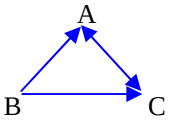
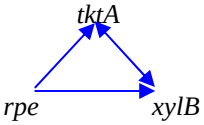
1. Motif 1.

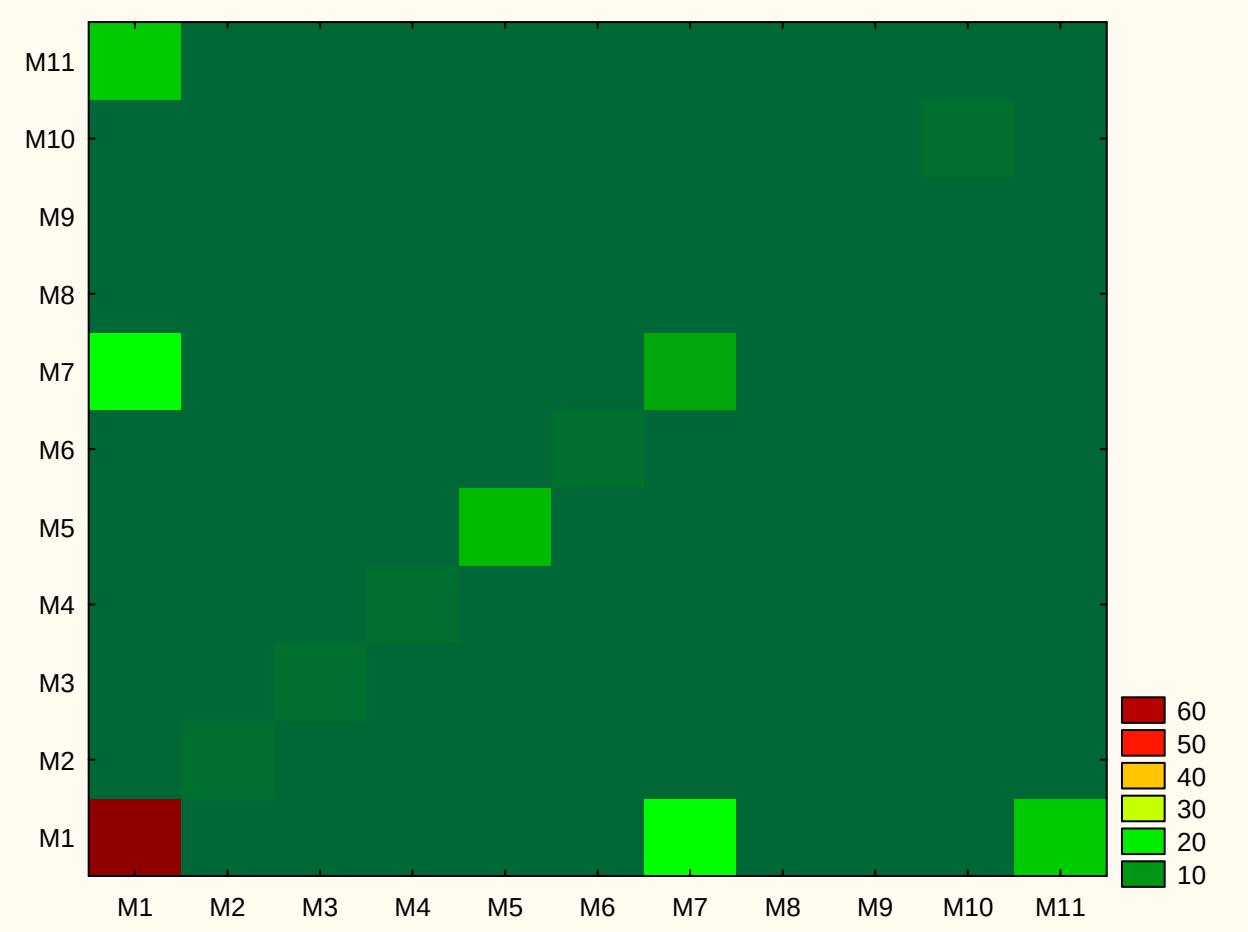
Motif pattern*	Motif Example	Zscore†	Nreal	RndAvg	RndStd
1. 		204.1	2220	92	10.4



A matrix representation of motif1 formed by the interaction of every specific pair of modules.

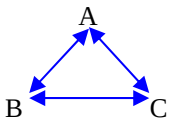
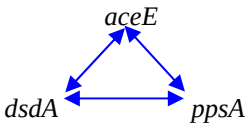
2. Motif 2.

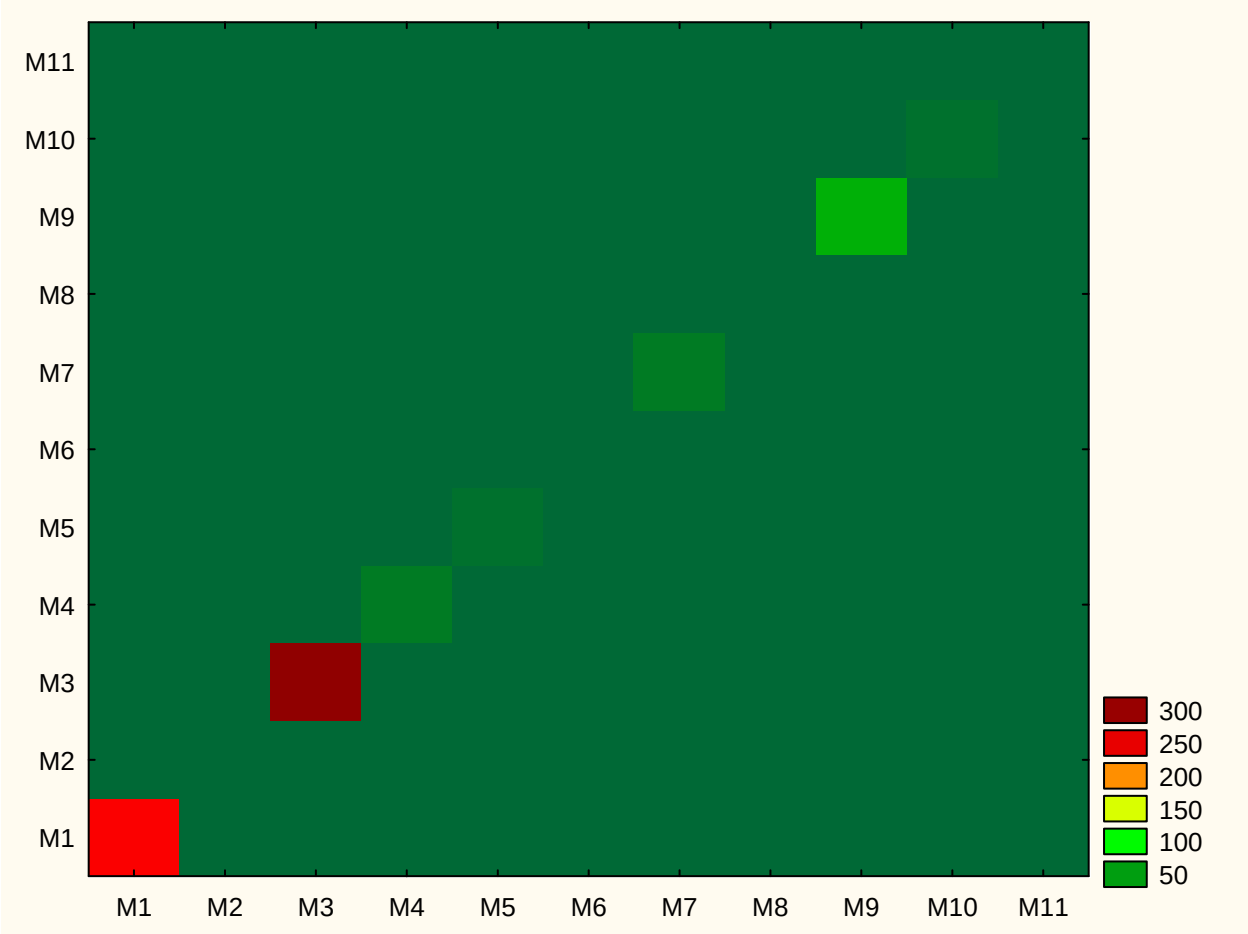
Motif pattern*	Motif Example	Zscore†	Nreal	RndAvg	RndStd
		179.7	625	10	3.4



A matrix representation of motif2 formed by the interaction of every specific pair of modules.

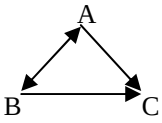
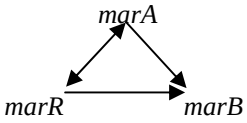
3. Motif 3.

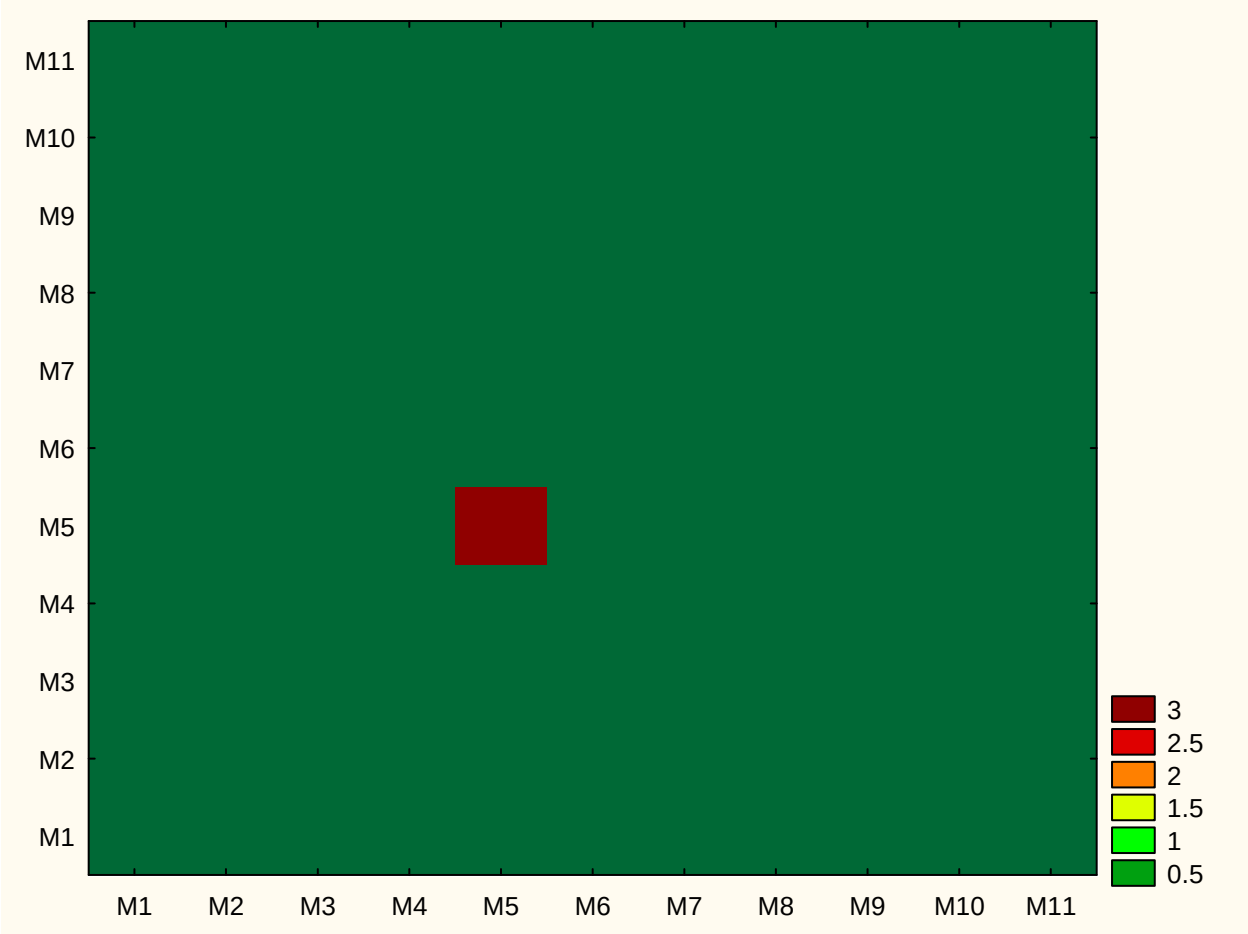
Motif pattern*	Motif Example	Zscore†	Nreal	RndAvg	RndStd
		176.4	3981	365	20.5



A matrix representation of motif3 formed by the interaction of every specific pair of modules.

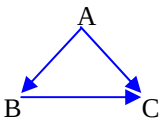
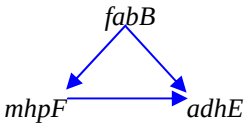
4. Motif 4.

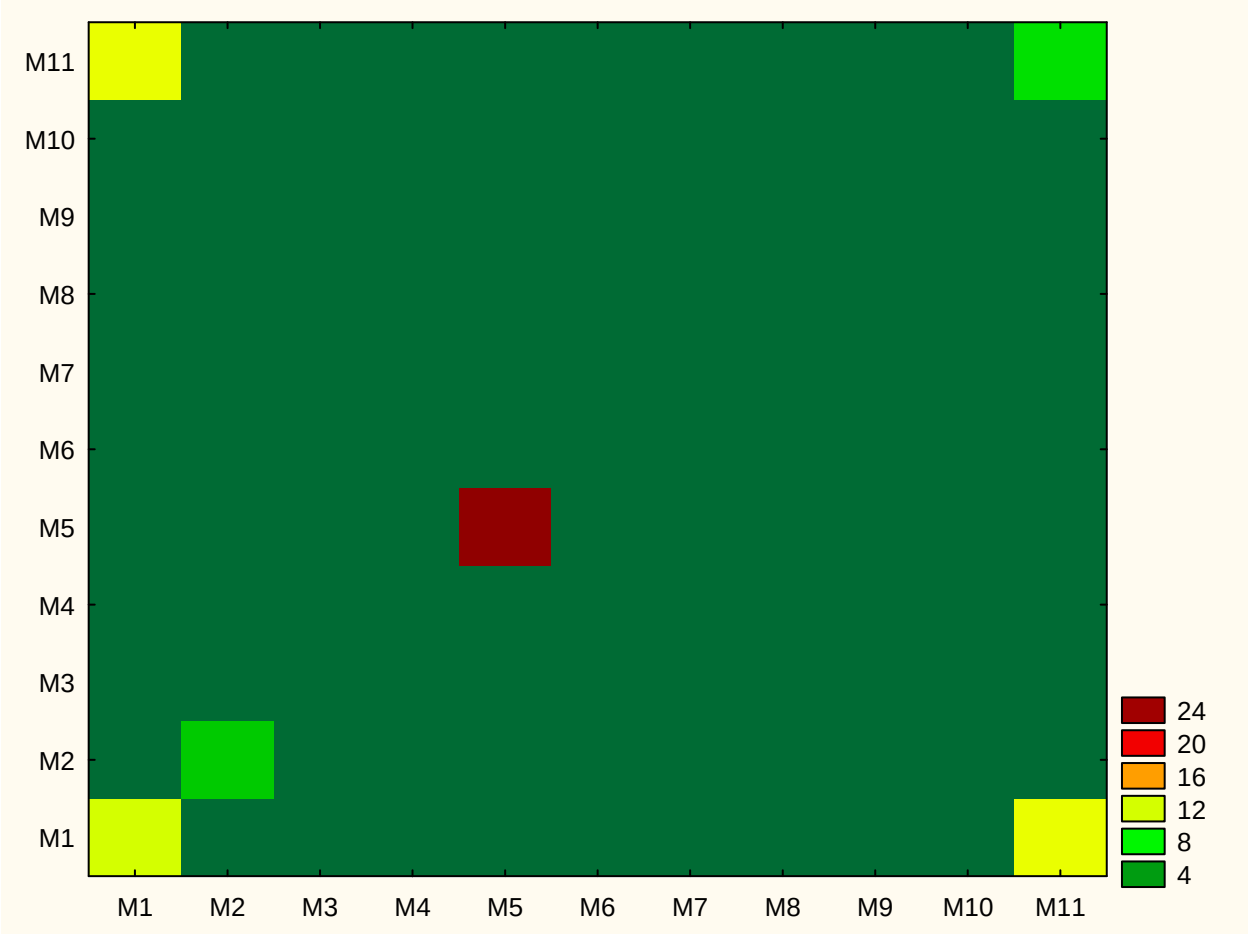
Motif pattern*	Motif Example	Zscore†	Nreal	RndAvg	RndStd
		25.7	50	2.1	1.9



A matrix representation of motif4 formed by the interaction of every specific pair of modules.

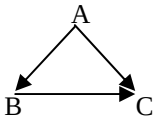
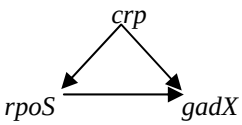
5. Motif 5.

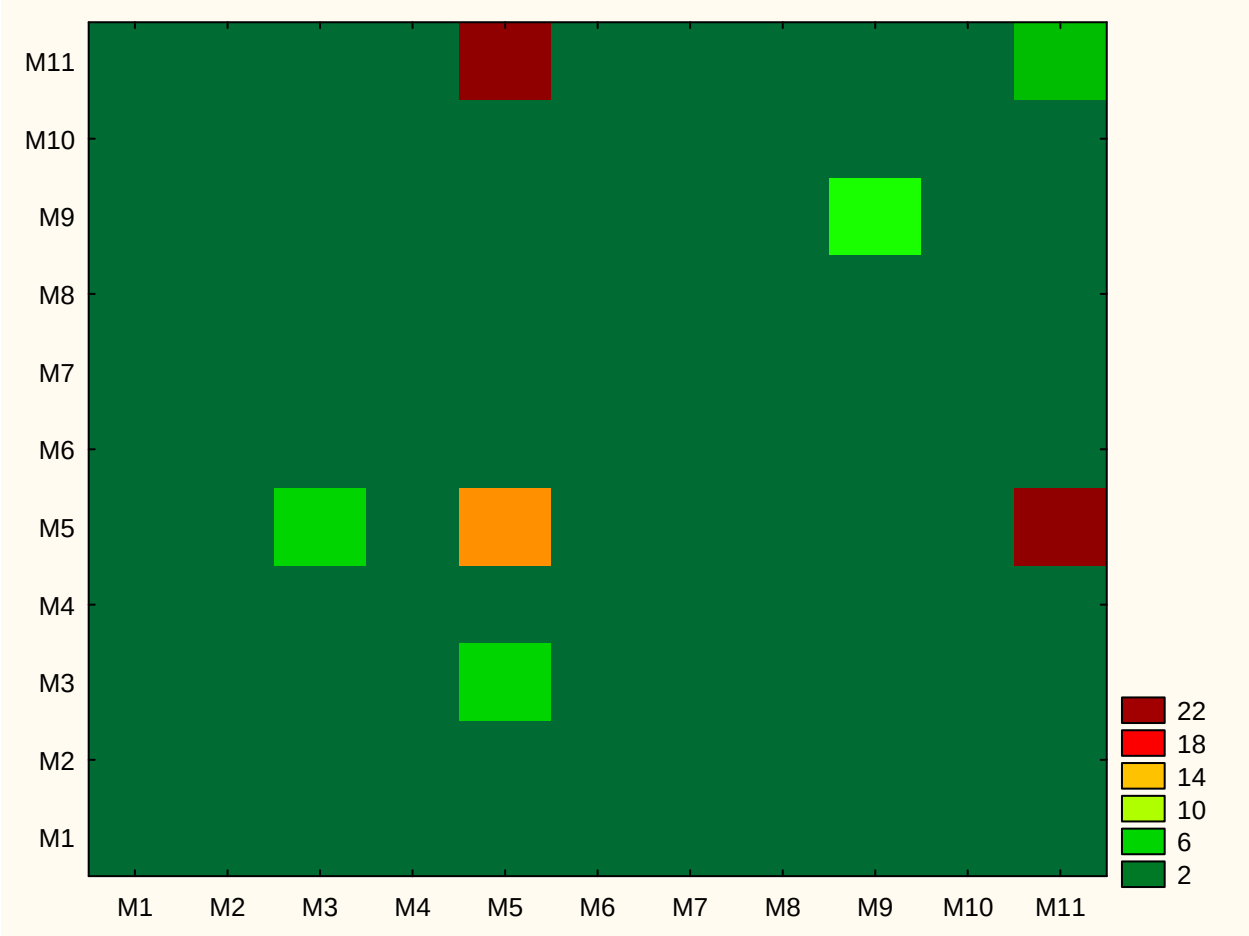
Motif pattern*	Motif Example	Zscore†	Nreal	RndAvg	RndStd
		24.8	344	90	10.2



A matrix representation of motif5 formed by the interaction of every specific pair of modules.

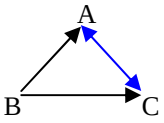
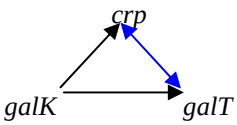
6. Motif 6.

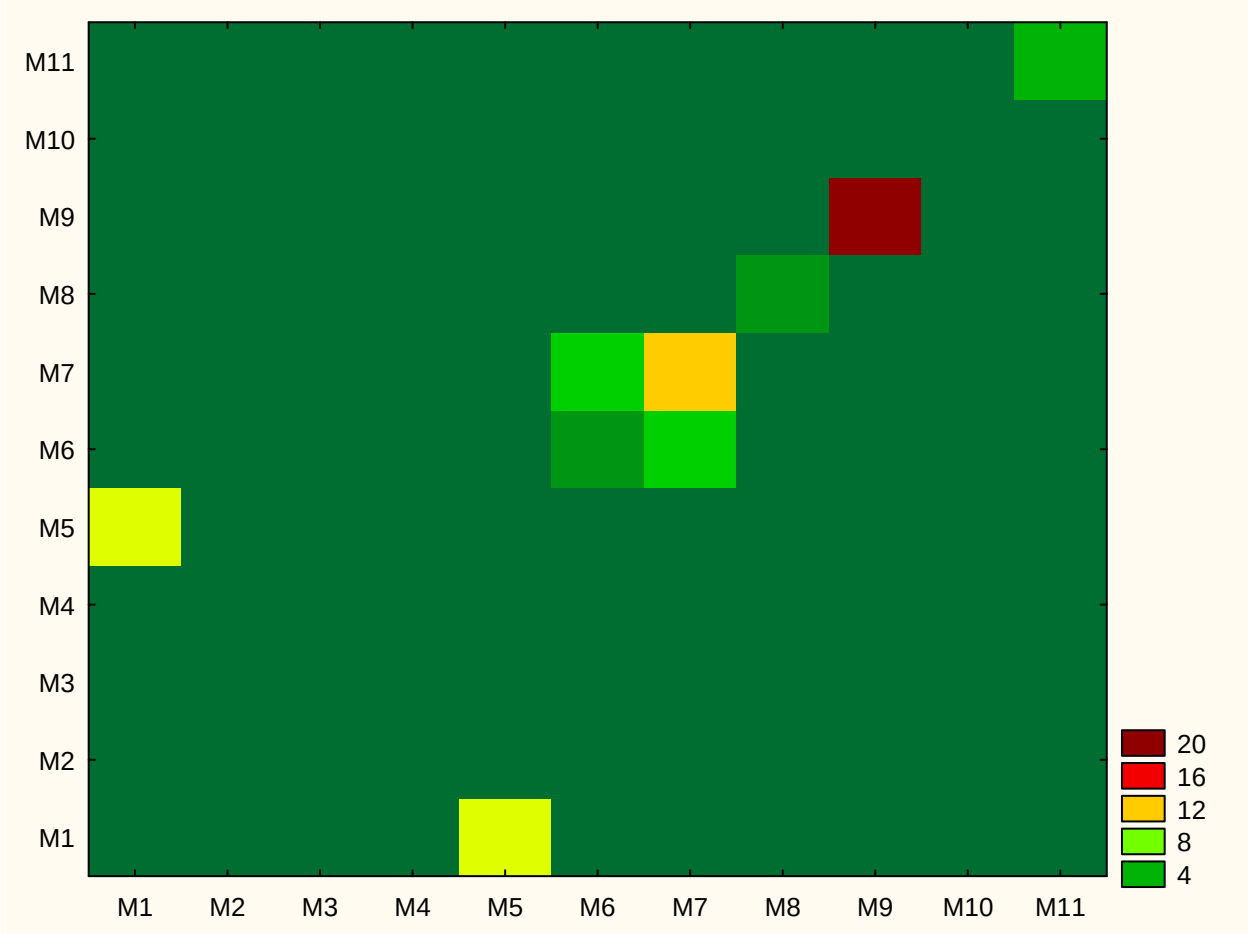
Motif pattern*	Motif Example	Zscore†	Nreal	RndAvg	RndStd
		12.4	285	86	16.1



A matrix representation of motif6 formed by the interaction of every specific pair of modules.

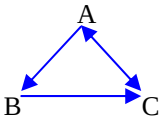
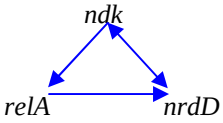
7. Motif 7.

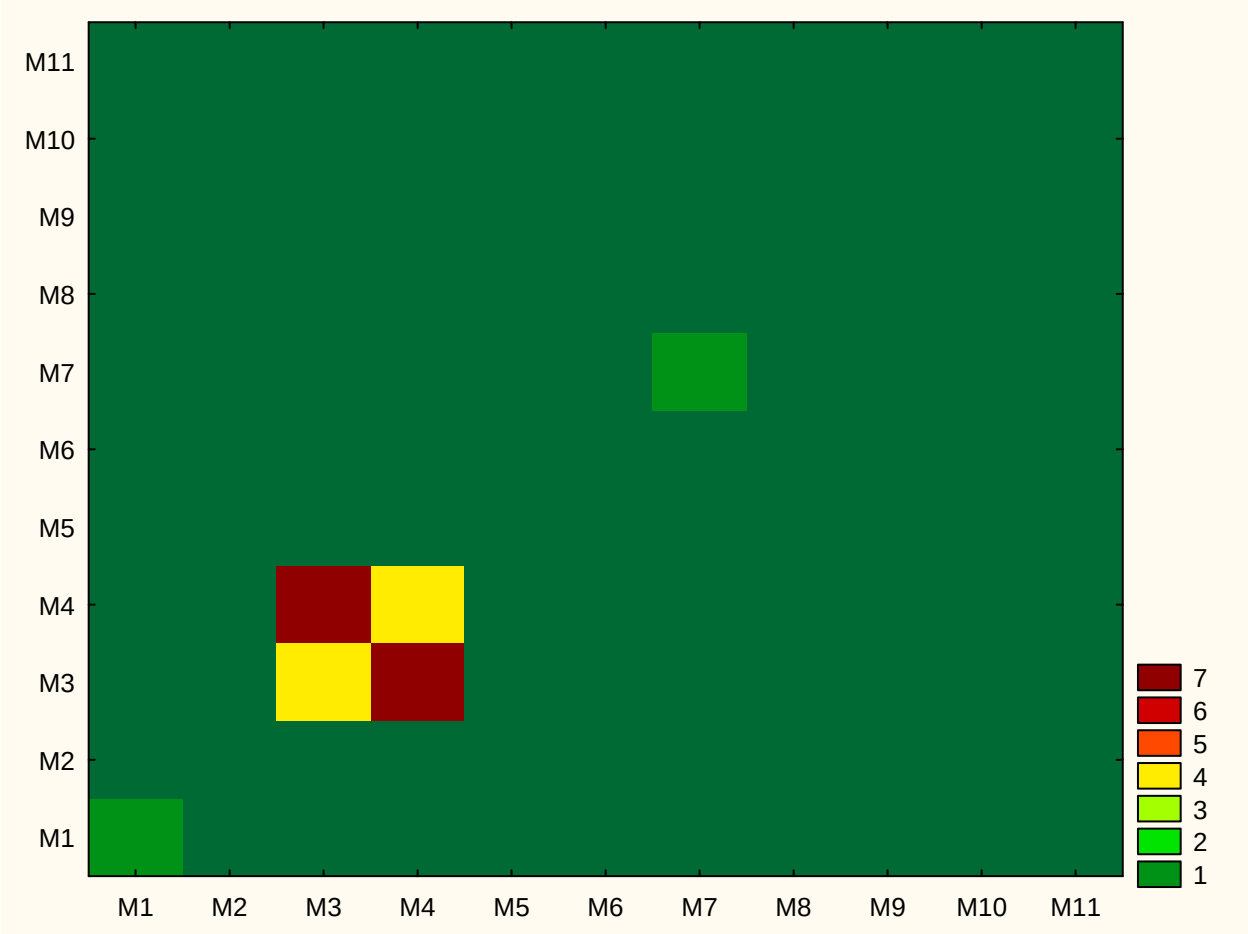
Motif pattern*	Motif Example	Zscore†	Nreal	RndAvg	RndStd
		11.5	280	107	15



A matrix representation of motif7 formed by the interaction of every specific pair of modules.

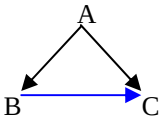
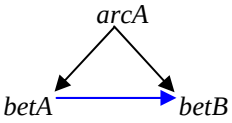
8. Motif 8.

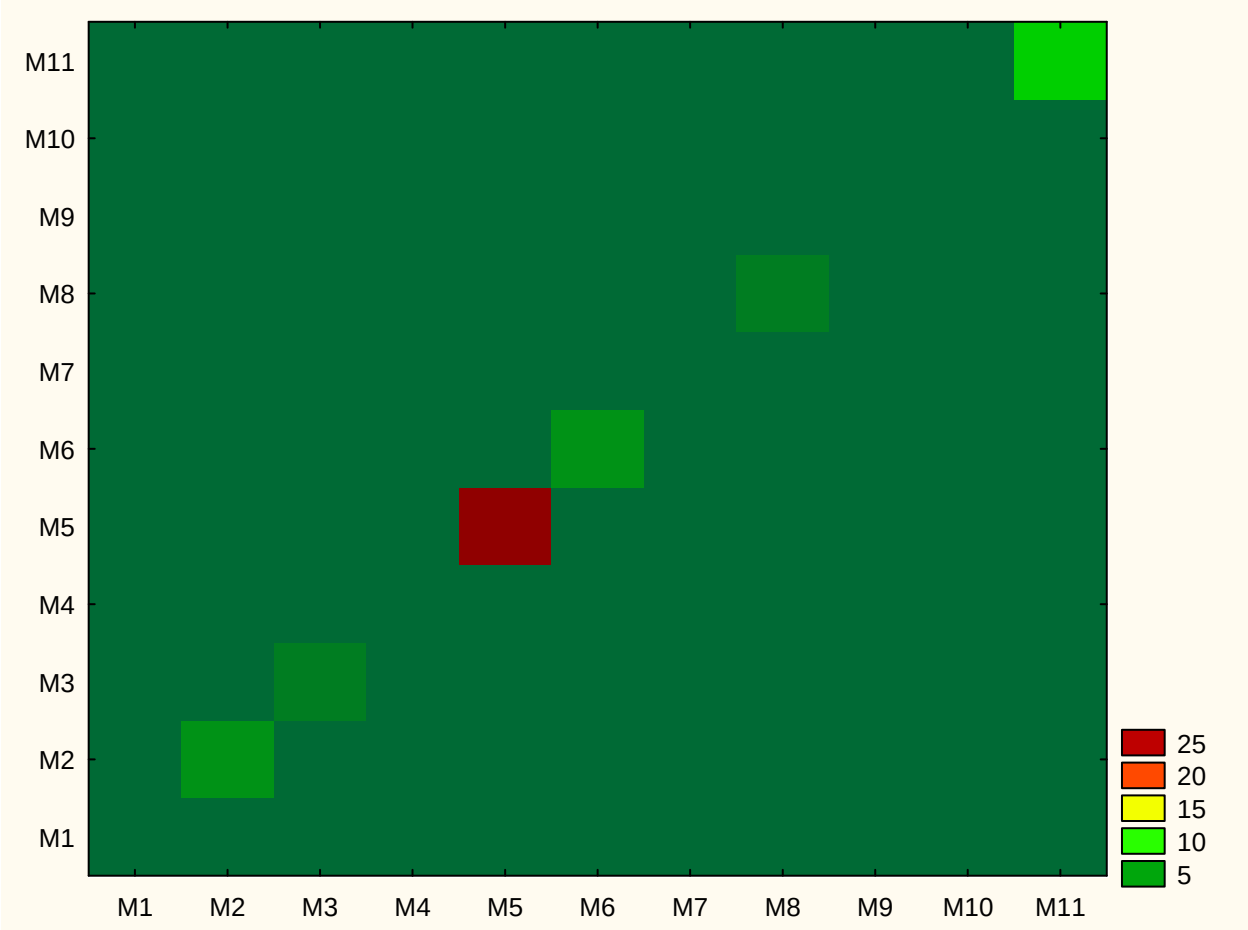
Motif pattern*	Motif Example	Zscore†	Nreal	RndAvg	RndStd
		11.4	100	31	6



A matrix representation of motif8 formed by the interaction of every specific pair of modules.

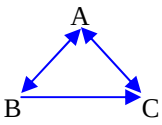
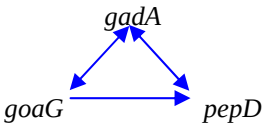
9. Motif 9.

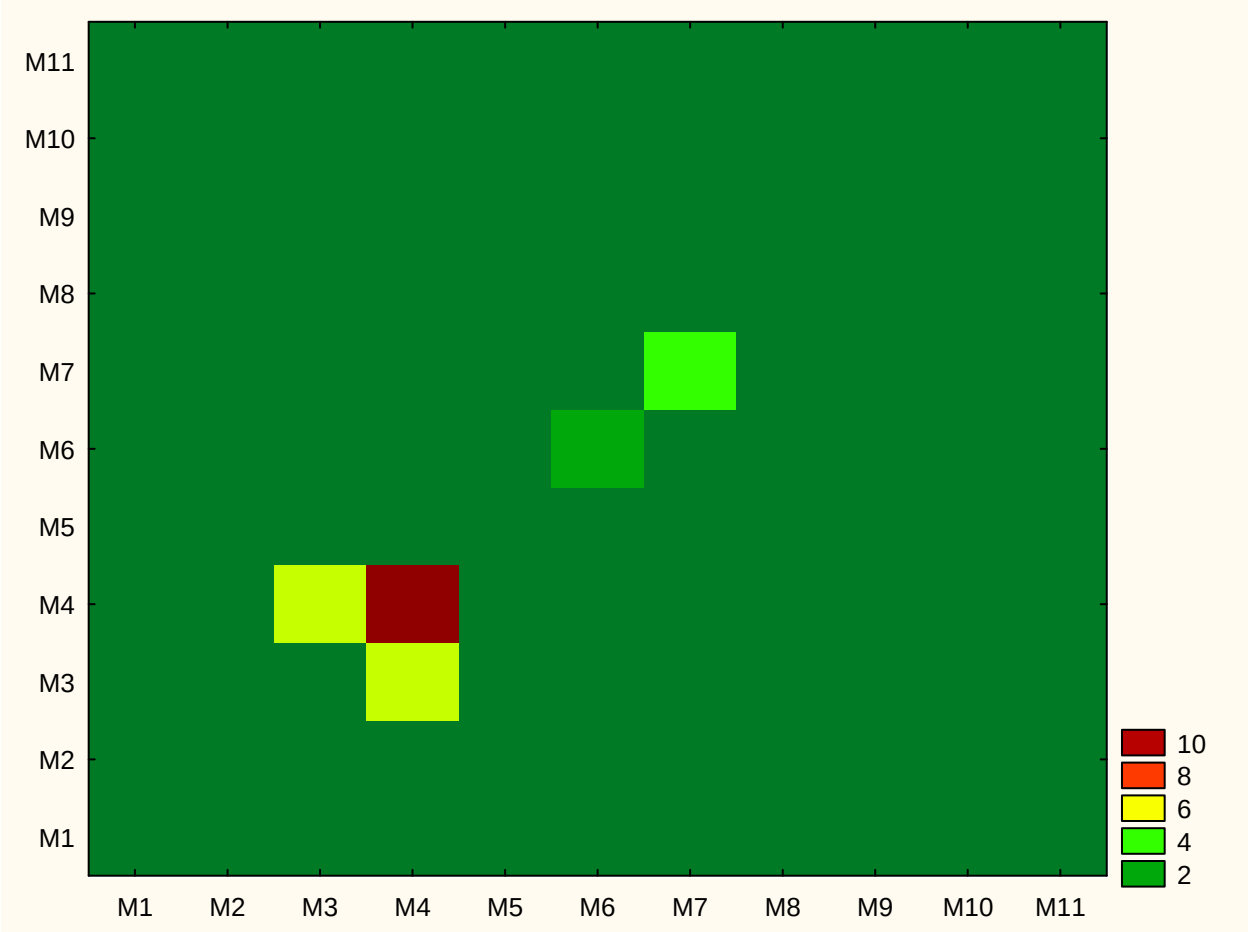
Motif pattern*	Motif Example	Zscore†	Nreal	RndAvg	RndStd
		10.8	219	92	11.8



A matrix representation of motif9 formed by the interaction of every specific pair of modules.

10. Motif 11.

Motif pattern*	Motif Example	Zscore†	Nreal	RndAvg	RndStd
		9.9	272	132	14.2



A matrix representation of motif11 formed by the interaction of every specific pair of modules.

References

Alkema, W. B., Lenhard, B. and Wasserman, W. W. (2004). "Regulog analysis: detection of conserved regulatory networks across bacteria: application to Staphylococcus aureus." *Genome Research* **14**: 1362-1373.

Auffray, C., Imbeaud, S., Roux-Rouquie´, M. and Hood, L. (2003). "From functional genomics to systems biology: concepts and practices." *CR Biologies* **326**: 879-892.

Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M. and Teichmann, S. A. (2004). "Distinct contributions of KH domains to substrate binding affinity of Drosophila P-element somatic inhibitor protein." *Curr. Opin. Struc. Biol.* **14**: 283-291.

Bairoch, A. (2000). "The ENZYME database in 2000." *Nucl. Acids Res.* **28**: 304-305.

Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S. and Young, R. A. e. a. (2003). "Computational discovery of gene modules and regulatory networks." *Nat Biotechnol.* **21**: 1337-1342.

Barabasi, A. L. and Oltvai, Z. N. (2004). "Network biology: Understanding the cell's functional organization." *Nat. Rev. Genet.* **5**: 101-113.

Batagelj, V. and Mrvar, A. (1998). "Pajek-a program for large network analysis." *Connections* **21**: 47-57.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C. and Eddy, S. R. (2004). "The Pfam protein families database." *Nucl. Acids Res.* **32**: 138-141.

Blattner, R., Plunkett, G., Bloch, A., Perna, T., Burland, Riley, Collado-Vides, J., Glasner, D., Rode, C. K. and Mayhew, G. F. e. a. (1997). "The complete genome sequence of *Escherichia coli K-12*." *Science* **277**: 1453-1474.

Bolouri, H. and Davidson, E. H. (2002). "Modeling Transcriptional regulatory networks." *BioEssays* **24**: 1118-1129.

Bouche, N. and Fromm, H. (2004). "GABA in plants: just a metabolite?" *Trends Plant Sci.* **9**: 110-115.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J. (2000). "Graph structure in the Web." *Computer Networks* **33**: 309-320.

Chuang, S. E., Daniels, D. L. and Blattner, F. R. (1993). "Global regulation of gene expression in *Escherichia coli*." *J. Bacteriol.* **175**: 2026-2036.

Collado-Vides, J., Magasanik, B. and Smith, T. F. (1996). "Integrative Approaches to Molecular Biology." *MIT Press, Cambridge Mass*.

Csete, M. E. and Doyle, J. C. (2004). "Bow ties, metabolism and disease." *Trends in Biotechnology* **22**: 446-450.

DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale." *Science* **278**: 680-686.

Dill, S., Kumar, R., McCurley, K., Rajagopalan, S., Sivakumar, D. and Tomkins, A. (2001). "Self-similarity in the web." *In Proceedings of the 27th VLDB Conference*: 69-78.

Dobrin, R., Beg, K., Barabasi, L. and Oltvai, N. (2004). "Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network." *BMC Bioinformatics* **5**: 10.

Donato, D., Leonardi, S., Millozzi, S. and Tsaparas, P. (2005). "Mining the inner structure of the web graph." *WebDB*: 145-150.

Girvan, M. and Newman, M. E. (2002). "Community structure in social and biological networks." *PNAS* **99**: 7821-7826.

Guelzim, N., Bottani, S., Bourgine, P. and Kepes, F. (2002). "Topological and causal structure of the yeast transcriptional regulatory network." *Nature Genetics* **31**: 60-63.

Gutierrez-Rios, R. M., Rosenblueth, D. A., Loza, J. A., Huerta, A. M., Glasner, J. D., Blattner, F. R. and Collado-Vides, J. (2003). "Regulatory network of Escherichia coli: consistency between literature knowledge and microarray profiles." *Genome Research* **13**: 2435-2443.

Herrgard, M. J., Covert, M. W. and Palsson, B. O. (2003). "Reconciling gene expression data with known genome-scale regulatory network structures." *Genome Research* **13**: 2423–2434.

Herrgard, M. J., Covert, M. W. and Palsson, B. O. (2004a). "Reconstruction of microbial transcriptional regulatory networks." *Curr. Opin.Biotechnol.* **15**: 70-77.

Herrgard, M. J. and Palsson, B. O. (2004b). "Flagellar biosynthesis in silico: building quantitative models of regulatory networks." *Cell* **117**: 689-690.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R. and L., H. (2001). "Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network." *Science* **292**: 929-934.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. and Barkai, N. (2002). "Revealing modular organization in the yeast transcriptional network." *Nature Genetics* **31**: 370-377.

Kanehisa, M. and Goto, S. (2000). "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucl. Acids Res.* **28**: 27-30.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M. (2005). "From genomics to chemical genomics: new
developments in KEGG." *Nucl. Acids Res.* **34**: D354-D357.

Kao, K. C., Yang, Y. L., Boscolo, R., Sabatti, C., Roychowdhury, V. and Liao, J. C. (2004). "Transcriptome-based determination of multiple transcription regulator activities in *Escherichia coli* by using network component analysis." *PNAS* **101**: 641-646.

Karp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., Tsoka, S., Darzentas, N., Kunin, V. and López-Bigas, N. (2005). "Expansion of the BioCyc collection of pathway/genome databases to 160 genomes." *Nucleic Acids Res.* **33**: 6083-6089.

Karp, P. D., Paley, S. and Romero, P. (2002b). "The Pathway Tools software." *Bioinformatics* **18**: S225-232.

Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Collado-Vides, J., Paley, S. M., Pellegrini-Toole, A., Bonavides, C. and Gama-Castro, S. (2002a). "The EcoCyc Database." *Nucl. Acids Res.* **30**: 56-58.

Kashtan, N., Itzkovitz, S., Milo, R. and Alon, U. (2004). "Topological generalizations of network motifs." *Phys. Rev. E.* **70**: 031909.

Keseler, I. M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I. T., Peralta-Gil, M. and Karp, P. D. (2005). "EcoCyc: a comprehensive database resource for *Escherichia coli*." *Nucl. Acids Res.* **33**: D334-D337.

Korbel, J. O., Jensen, L. J. and Von Mering, C., et al. (2004). "Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs." *Nat Biotechnol.* **22**: 911-917.

Krieger, C. J., Zhang, P., Mueller, L. A., Wang, A., Paley, S., Arnaud, M., Pick, J., Rhee, S. Y. and Karp, P. D. (2004). "MetaCyc: a multiorganism database of metabolic pathways and enzymes." *Nucl. Acids Res.* **32**: D438–D442.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J. B., Volkert, T. L.,

Fraenkel, E., Gifford, D. K. and Young, R. A. (2002). "Transcriptional regulatory networks in *Saccharomyces cerevisiae*." *Science* **298**(5594): 799-804.

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A. and Gerstein, M. (2004). "Genomic analysis of regulatory network dynamics reveals large topological changes." *Nature* **431**: 308-312.

Ma, H. W., Kumar, B., Ditges, U., Gunzer, F., Buer, J. and Zeng, A. P. (2004c). "An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs." *Nucl. Acids Res.* **32**: 6643-6649.

Ma, H. W. and Zeng, A. P. (2003a). "Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms." *Bioinformatics* **19**: 270-277.

Ma, H. W. and Zeng, A. P. (2003b). "The connectivity structure, giant strong component and centrality of metabolic networks." *Bioinformatics* **19**: 1423-1430.

Ma, H. W., Zhao, X. M., Yuan, Y. J. and Zeng, A. P. (2004a). "Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph." *Bioinformatics* **20**: 1870-1876.

Ma, W., Buer, J. and Zeng, P. (2004b). "Hierarchical structure and modules in the *Escherichia coli* transcriptional regulatory network revealed by a new top-down approach." *BMC Bioinformatics* **5**: 199.

Maeta, K., Izawa, S., Okazaki, S., Kuge, S. and Inoue, Y. (2004). "Activity of the Yap1 transcription factor in *Saccharomyces cerevisiae* is modulated by methylglyoxal, a metabolite derived from glycolysis." *Mol. Cell. Biol.* **24**: 8753-8764.

Mangan, S. and Alon, U. (2003a). "Structure and function of the feed-forward loop network motif." *PNAS* **100**: 11980-11985.

Mangan, S., Zaslaver, A. and Alon, U. (2003b). "The coherent feed forward loop serves as a sign-sensitive delay element in transcription networks." *J. Mol. Biol.* **334**: 197-204.

Martinez-Antonio, A. and Collado-Vides, J. (2003). "Identifying global regulators in transcriptional regulatory networks in bacteria." *Curr. Opin. Microbiol.* **6**: 482-489.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E. and Kel-Margoulis, O. V. e. a. (2003). "TRANSFAC: transcriptional regulation, from patterns to profiles." *Nucl. Acids Res.* **31**: 374-378.

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. and Alon, U. (2004). "Superfamilies of Evolved and Designed Networks." *Science* **303**: 1538-1542.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002). "Network Motifs: Simple Building Blocks of Complex Networks." *Science* **298**(5594): 824-827.

Nunn, W. D., Kelly, D. L. and Stumfall, M. Y. (1977). "Regulation of fatty acid synthesis during the cessation of phospholipid biosynthesis in *Escherichia coli*." *J. Bacteriol.* **132**(526-531).

O' Malley, M. A. and Dupre´, J. (2005). "Fundamental issues in systems biology." *BioEssays* **27**: 1270-1276.

Palsson, B. O. (1997). "What lies beyond bioinformatics?" *Nat Biotechnol.* **15**: 3-4.

Papin, J. A., Price, N. D. and Palsson, B. O. (2002). "Extreme Pathway Lengths and Reaction Participation in Genome-Scale Metabolic Networks." *Genome Research* **12**: 1889-1900.

Papin, J. A., Stelling, J., Price, N. D., Klamt, S., Schuster, S. and Palsson, B. O. (2004). "Comparison of network-based pathway analysis methods." *Trends in Microbiology* **22**: 400-405.

Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J. and Giles, C. L. (2002). "Winners don't take all: Characterizing the competition for links on the web." *PNAS* **99**: 5207-5211.

Price, N. D., Reed, J. L., Papin, J. A., Wiback, S. J. and Palsson, B. O. (2003). "Network-based analysis of metabolic regulation in the human red blood cell." *Journal of Theoretical Biology* **225**: 185-194.

Ranea, J. A., Buchan, D. W., Thornton, J. M. and Orengo, C. A. (2004). "Evolution of protein superfamilies and bacterial genome size." *J. Mol. Biol.* **336**: 871-887.

Ravasz, E., Somera, A., Mongru, A., Oltvai, Z. and Barabasi, A. (2002). "Hierarchical Organization of Modularity in Metabolic Networks." *Science* **297**: 1551-1555.

Resendis-Antonio, O., Freyre-González, J. A., Menchaca-Méndez, R., Gutiérrez-Ríos, R. M., Martínez-Antonio, A., Cristhian Ávila, C. and Collado-Vides, J. (2005). "Modular analysis of the transcriptional regulatory network of *E. coli*." *Trends Genet.* **21**: 16-20.

Riley, M. (1998). "Genes and proteins of *Escherichia coli K-12*." *Nucl. Acids Res.* **26**(1): 54.

Rock, C. O. and Jackowski, S. (1982). "Regulation of Phospholipid Synthesis in *Escherichia coli*." *J. Biol. Chem.*: 10759-10765.

Rudd, E. (2000). "EcoGene: a genome sequence database for *Escherichia coli K-12*." *Nucl. Acids Res.* **28**: 60-64.

Salgado, H., Gama-Castro, S., Martinez-Antonio, A., Diaz-Peredo, E., Sanchez-Solano, F., Peralta-Gil, M., Garcia-Alonso, D., Jimenez-Jacinto, V., Santos-Zavaleta, A. and Bonavides-Martinez, C. e. a. (2004). "RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli K-12*." *Nucl. Acids Res.* **32**(Database): 303-306.

Schomburg, I., Chang, A., Hofmann, O., C., E., F., E. and D., S. (2002b). "BRENDA: a resource for enzyme data and metabolic information." *Trends Biochem Sci.* **27**(1): 54-56.

Schomburg, I., Chang, A. and Schomburg, D. (2002a). "BRENDA, enzyme data and metabolic information." *Nucl. Acids Res.* **30**: 47-49.

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G. and Schomburg D. (2004). "BRENDA, the enzyme database: updates and major new developments." ***Nucl. Acids Res.*** 1;32 Database issue:D431-3.

Schuster, S., Fell, D. A. and Dandekar, T. (2000). "A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks." *Nat Biotechnol.* **18**: 326-332.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. and Friedman, N. (2003). "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data." *Nature Genetics* **34**: 166-176.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003). "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research* **13**: 2498-2504.

Shen-Orr, S. S., Milo, R., Mangan, S. and Alon, U. (2002). "Network motifs in the transcriptional regulation network of *Escherichia coli*." *Nature Genetics* **31**: 64-68.

Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S. and Gilles, E. D. (2002). "Metabolic network structure determines key aspects of functionality and regulation." *Nature* **420**: 190-193.

Strohman, R. C. (1997). "The coming Kuhnian revolution in biology." *Nat Biotechnol.* **15**: 194-200.

Uetz, P., Ideker, T. and Schwikowski, B. (2002). "Visualization and integration of protein-protein interactions." *Cold Spring Harbor Laboratory Press*: 623-646.

Van Nimwegen, E. (2003). "Scaling laws in the functional content of genomes." *Trends Genet.* **19**: 479-484.

VanBogelen, R. A., Abshire, K. Z., Pertsemlidis, A., Clark, R. L. and Neidhardt, F. C. (1996). " In Cellular and Molecular Biology: Escherichia coli and Salmonella, Geneprotein database of *Escherichia coli K-12*." *Washington, D.C. ASM* **Edition 6**.

Wall, M. E., Hlavacek, W. S. and Savageau, M. A. (2004). "Design of gene circuits: lessons from bacteria." *Nature Rev. Genet.* **5**: 34-42.

Wang, M. X. and Church, G. M. (1992). "A whole-genome approach to in vivo DNA-protein interactions in *E.coli*." *Nature* **360**: 606-610.

Warren, P. B. and Wolde, P. R. T. (2004). " Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of *Escherichia coli*." *J Mol Biol.* **342**: 1379-1390.

Wolf, M. and Arkin, P. (2003). "Motifs, modules and games in bacteria." *Curr. Opin. Microbiol.* **6**: 125-134.

Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R. Y., Alon, U. and Margalit, H. (2004). "Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction." *PNAS* **101**: 5934-5939.

Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J. D., Bertin, N., Chung, S., Vidal, M. and Gerstein, M. (2004). "Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs." *Genome Research* **14**: 1107-1118.

Yu, H., Luscombe, N. M., Qian, J. and Gerstein, M. (2003). "Genomic analysis of gene expression relationships in transcriptional regulatory networks." *Trends Genet.* **19**: 422-427.

Zhang, L. V., King, O. D., Wong, S. L., Goldberg, D. S., Tong, H. Y. A., Lesage, G., Andrews, B., Bussey, H., Boone, C. and Roth, F. P. (2005). "Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network." *Journal of Biology* **4**: 6.

Lebenslauf

BHARANI KUMAR

EDUCATION

PhD (Systems Biology, 2004-2007)
Helmholtz Zentrum für Infektionsforschung,
Braunschweig, Germany

Postgraduate Course in Bioinformatics (2003-2004)
CUBIC, University of Köln,
Köln, Germany

Bachelors of Technology (Chemical Engineering) (1996-2000)
Coimbatore Institute of Technology,
Coimbatore, India

PROJECT

PhD Work

TITLE 1 – **Vertical integration and global analysis of molecular interactions network of *E.coli***

Description:

Phenotypical characteristics of cells often arise from interactions between genes, proteins and metabolites. For a complete understanding of cellular processes and their regulations it is necessary to vertically integrate the molecular networks into an interactome and understand its global structure. Here, we constructed an integrated molecular network (IMN) of *E. coli* from metabolic reactions, metabolite-protein interactions (MPI) and transcriptional regulation data. We studied three fundamental aspects of cellular processes: (i) feedback regulation of gene expression, (ii) network motifs and (iii) global organization. Intriguingly, we found that feedback regulation of gene expression in *E. coli* is exclusively mediated by MPIs and 69 such feedback loops (FBLs) were identified. Further analysis of IMN revealed a bow-

tie connectivity structure spanning three molecular levels in a nested way. All the FBLs were only found in the core of the bow-tie. We detected thirteen three-node network motifs comprising five composite motifs which included at least two different types of interactions and analyzed their significance in the bow-tie. About 75% of them are interconnected, thereby forming the backbone of GSC. This study demonstrated the necessity and usefulness of a vertical integration of molecular networks and its global analysis for better understanding cellular processes and their regulation.

Availability : The integrated molecular network database and feedback regulations through MPI is available in excel format.

TITLE 2 – **Modularity analysis of integrated molecular networks of *E.coli***

Description:

Emergent interactions and dependencies between disparate molecules (e.g. genes, proteins or metabolites) make the biological system "complex". A major challenge is to develop the most appropriate, fast and robust method for clustering such a heterogeneous molecular circuitry of interactions. Here, we have used modularity as a parameter of clustering and performed modular analysis on an integrated molecular network of Escherichia coli constructed from diverse collection of datasets involving metabolic reactions, metabolite protein interactions and transcriptional regulation. We have found that clustering this complex network significantly grouped together genes of known similar function in well-defined physiologically related modules. Identification of network motifs and correlating them with the modules of highly connected nodes may define their potential functional role. To this end, we detected and analyzed twelve highly significant three-node network motifs among which four are composite network motifs comprising multiple types of interactions. Distribution analysis of these motifs within and between the various functional modules supported the fact that these motifs represent basic patterns of regulation and organization of genes into modules. This study presents a basic framework for detecting functional modules and their interaction with various motifs in an integrated E.coli system.

Availability: The version of the cluster tool is distributed as a package.

Postgraduate Level

PROJECT TITLE - Development and application of a consensus scoring function for Protein-protein docking

Description:

In the workgroup of Prof.Schomburg a protein-protein docking software is being developed since 1996. A Fast Fourier Transform (FFT) - based docking algorithm creates a list of possible complex conformations from the two individual input structures, based on geometric correlation. The software is now second generation and a variety of scoring functions has been added to the pure geometrical docking procedure. All these scoring functions attempt to rank the proposed complexes so that those structures closest to the native complex can possibly be recognized from the large number of generated false positives. Currently all scoring functions work independently, each establishing its own ranking based on various properties like electrostatics, hydrophobicity, knowledge based atom-pair-potential etc.

Project:

All these individual scoring functions are combined to a consensus scoring function i.e. by simply calculating a normalized weighted sum of all individual scoring schemes. The weighting factors are then optimized by training a Neural Network & other machine learning algorithm (SVM, Bayesian Network). There are vital properties which are integrated that have not been implemented so far (i.e. desolvation) as well as the addition of a final minimization step for the highest ranking structures.

PROJECT TITLE - Molecular evolution of protein atomic composition in Cyanobacteria

Objective - Demonstrate the systematic occurrence of atomic bases in assimilatory proteins of Cyanobacteria.

Undergraduate Level

PROJECT TITLE - Manufacture of 10 T.P.D. OF Ethanol by molasses fermentation

Objective - The study was aimed at manufacture of 10 T.P.D. of ethanol by fermentation of molasses, designing of major equipments- Absorber Distillation column, Heat Exchanger: Cost Estimation & Project Feasibility.

CERTIFICATION CREDENTIALS

Sun Certified Programmer for Java 2 Platform
(Candidate ID: 100089014, Reg.No: H26SYD002B)

Brain bench Certified –Java 2.0
(Transcript ID-2466863)

Graduate Record Examinations (GRE)-1630 Points

TOEFL (Computer-based Test)-270 Points

TECHNICAL SKILLS

Operating Systems - DOS, LINUX & Windows

Language(s) - Java 2, C, C++

Databases - MS-Access, Oracle 8i

Web Technologies - HTML, Applets, Swings, JDBC, JSP,Servlets, Java Beans, XML, EJB with Web logic Server.
